Title

Optical emissivity dataset of multi-material heterogeneous designs generated with automated figure extraction

Journal

Scientific Data, 9(1)

ISSN

2052-4463

Authors

Baibakova, Viktoriia

Elzouka, Mahmoud

Lubner, Sean

et al.

Publication Date

2022

DOI

10.1038/s41597-022-01699-3

scientific data

OPEN

DATA DESCRIPTOR

Optical emissivity dataset of multi-material heterogeneous designs generated with automated figure extraction

Viktoriia Baibakova, Mahmoud Elzouka, Sean Lubner, Ravi Prasher & Anubhav Jain ✉

Optical device design is typically an iterative optimization process based on a good initial guess from prior reports. Optical properties databases are useful in this process but difficult to compile because their parsing requires finding relevant papers and manually converting graphical emissivity curves to data tables. Here, we present two contributions: one is a dataset of thermal emissivity records with design-related parameters, and the other is a software tool for automated colored curve data extraction from scientific plots. We manually collected 64 papers with 176 figures reporting thermal emissivity and automatically retrieved 153 colored curve data records. The automated figure analysis software pipeline uses Faster R-CNN for axes and legend object detection, EasyOCR for axes numbering recognition, and k-means clustering for colored curve retrieval. Additionally, we manually extracted geometry, materials, and method information from the text to add necessary metadata to each emissivity curve. Finally, we analyzed the dataset to determine the dominant classes of emissivity curves and determine the underlying design parameters leading to a type of emissivity profile.

Background & Summary

Optical device design has impacted many fields, from the pioneering work of Fritts on the selenium solar cell[1] to the cutting-edge elaboration of nanophotonic intercellular force sensors expanding the conventional microbiology toolkit[2]. Further progress became possible due to materials synthesis[3,4] and modeling[5,6] advancements, allowing precise light manipulation over a wide range of wavelengths. Nevertheless, device design optimization remains an iterative process, strongly relying on a good initial guess followed by potentially time-consuming optimization. Modern sources of successful and useful initial designs are databases compiled from digesting the relevant literature, such as Materials Platform for Data Science[7] and HITRAN[8].

Optical properties databases should cover as many materials as possible and be up-to-date. There have been several notable endeavors[9–13] to translate literature into structured databases by parsing the text. However, text-based parsing of data is insufficient for many material properties because much of the needed information is communicated through graphs (e.g., spectral data). The standard method[14] for converting graphs is manual curve extraction using software such as WebPlotDigitizer[15], MATLAB GRABIT[16], DataThief[17]. Manual extraction requires significant user participation (i.e., clicking along the curve). In our experience, it takes approximately 3 minutes to parse a simple graph, which is practical for small tasks but becomes limiting if hundreds of graphs must be extracted. In contrast, existing efforts to automate graph data extraction have a list of drawbacks, such as parsing only continuous curves without sharp picks[18], requiring the figures to have PDF embedded axes[19] or having incompatibility issues due to no longer being actively maintained[20]. Therefore, the need exists for a tool for automated curve extraction from plots.

To address the listed issues, we compiled a dataset of thermal emissivity measurements from the optical scientific literature using various image analysis techniques. Figure 1 reviews the overall pipeline, which includes the following steps. First, we manually collected a corpus of 64 relevant publications. From these, we manually retrieved 176 figures containing emissivity-wavelength data relations. We implemented an algorithm for automated curve raw data extraction and automatically obtained data records for 153 curves. Next, we manually extracted two types of metadata from the text: the general information on the publication and design-related

Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley, CA, 94720, USA. ✉e-mail: ajain@lbl.gov

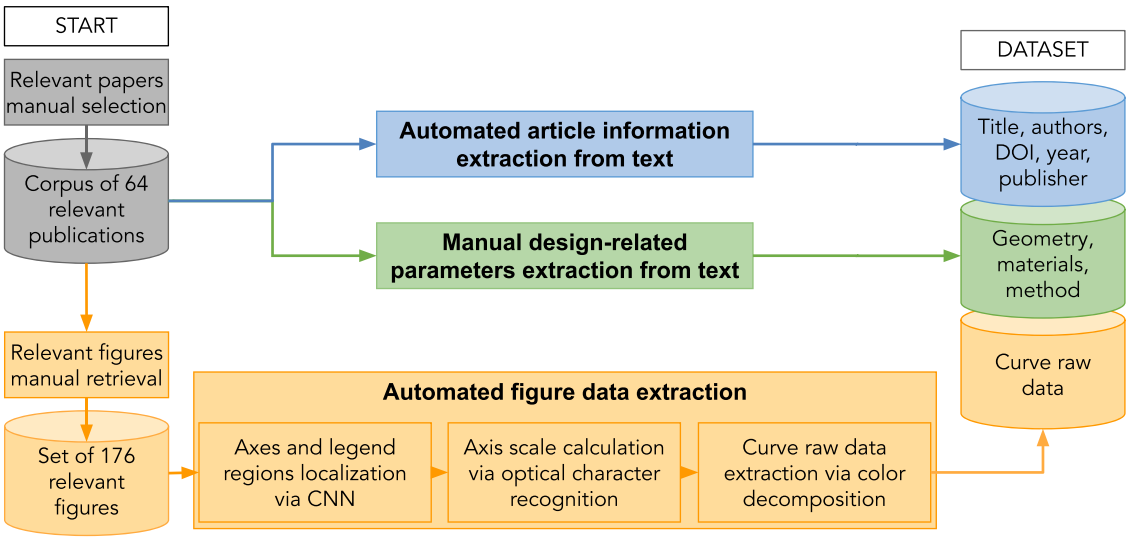

Fig. 1 The overall pipeline of data collection and organization into dataset. The data is retrieved from a corpus of 64 manually collected relevant papers (gray). There are three categories of data retrieval: blue - automated extraction of the general article information from text; green - manual extraction of the design-related parameters; and orange - automated curve raw data extraction embedded in semi-automatic figure analysis.

parameters such as the materials used and device geometry corresponding to each curve. Finally, we wrapped the collected information into an explicit dataset record.

This paper presents a dataset of thermal emissivity of multi-material heterogeneous designs and the algorithm used for its creation. Regarding the dataset, this work describes the chosen data format and dataset organization. It also covers the technical validation of the collected records and provides a use case for the dataset. For the algorithm, the article presents a detailed description of the method used to collect every data entry. It addresses the aggregation of general information like DOI, publisher, authors, year, and title. Also, it covers the retrieval of materials, design, and method descriptions. Additionally, the paper reports the performance of a proposed tool for automated curve extraction.

Methods

We established an algorithm that automates data extraction from figures and produced a comprehensive dataset of optical properties. Figure 1 provides an overview of the complete workflow; the various steps are described next in greater detail.

Generation of initial corpus. We manually collected the corpus of 64 publications[21–84] referring to emissivity by keeping track of relevant articles during our routine research for several years. We further used Google Scholar to search for articles by keywords and extracted more papers from the references. All selected articles contained graphical information of interest: emissivity-wavelength dependencies depicted as 2D curves on a blank background.

Automated article information extraction from text. General information on publications (blue path and dataset component on Fig. 1) was extracted automatically using Mendeley[85]. We saved the corpus as a Mendeley archive, which allowed us to export it as a single BibTeX[86] file containing the desired information. From the formatted BibTeX files, we used regular expressions[87] to retrieve the DOI, title, authors, publisher, URL, and year of publication for each article. We note that dedicated software libraries for parsing BibTeX files such as pybtex[88] (Python) are also available; however, we did not use those in this work.

Manual design-related parameters extraction from text. Design-related parameters (green path and dataset component on Fig. 1) are commonly reported in different sections of a publication, making it challenging to connect each curve record with its corresponding device geometry (sandwich, thin-film, grating), list of materials (W, Al, SiC), and method of data generation (calculation, experiment). Figure 2 demonstrates that design-related parameters of a given dataset record (data curve) were included in unrelated snippets of a sample paper. In the example from Fig. 2, the emissivity figure caption contains an incomplete list of materials and a brief geometry description elaborated in the figure-referring text. However, the full description of these two parameters is only given in the synthesis section of the paper, which already refers to other figures and does not mention the one with emissivity curves. We also note some complicated cases[49] in our corpus when the authors reported the design solely graphically, leaving out the explanation in the text. Regarding the method of data generation, it could be reported in any location throughout the paper, and while it was usually possible to distinguish between experiment and theory from the context, many authors[23,34] did not completely specify the used tool. For the issues listed above and others, our attempts to develop automatic tools for metadata extraction were insufficient to obtain the desired attributes (see SI.1), and this analysis was conducted through a manual approach.

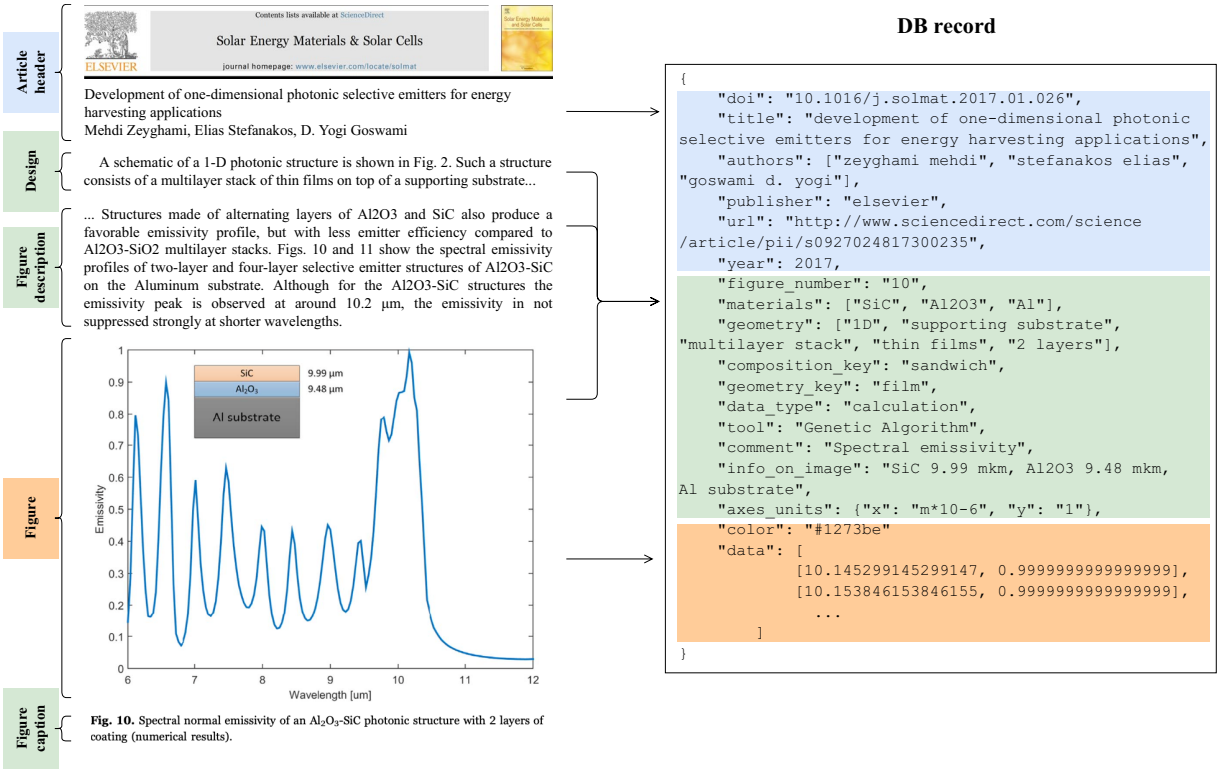

Fig. 2 Information extraction from the source paper to the dataset record. Colors correspond to the category of extraction: blue - automated text analysis; green - manual text analysis; orange - automated figure analysis. Information is taken from different parts of an article; for example, materials are listed in the figure description and within the figure itself. This example uses the work of Zeyghami *et al.*[24].

We manually located text passages containing information about each curve and recorded this information. First, we recorded all distinct materials (chemical compositions) used in the device. We categorized records into two groups: "single material" if the structure was made out of a single material or "sandwich" if the structure was a multilayer design. We also parsed all keywords related to the device geometry. We found 100 distinct descriptors (thin film, aperiodic multilayers, 2D array, front coating). Using them, we classified geometry descriptions into seven types. These were: (i) film, (ii) 1D grating - a film with an array of slots of any shape on the surface with 1-dimensional periodicity (with or without coating), (iii) 2D grating - a film with an array of slots of any shape on the surface with 2-dimensional periodicity (with or without coating), (iv) 2D cylindrical cavities - a film with an array of cylindrical holes on the surface with 2-dimensional periodicity (empty or filled), (v) wire, (vi) bull's eye - a film with a concentric equally spaced circular grooves on the surface (sometimes also with coating), and (vii) microspheres - random media composed of microscopic balls. Lastly, we parsed the method of data generation: experiment or computation and the characterization or modeling tool (Fourier transform infrared spectroscopy, finite-difference time-domain, etc.).

Figure data extraction. The next step of the procedure was to detect emissivity records from graphs and parse them (orange regions on Fig. 2). We examined 64 papers for the graphical information of interest: emissivity-wavelength dependencies depicted as 2D curves on a blank background. We found 176 images with 550 thermal emissivity curves and manually converted them to PNG format. We manually split figures with multiple plots for the final one to contain a single plot panel and axis with units. The figures varied from 600 to 1400 pixels in width and 800 to 2000 pixels in height.

We followed a three-step algorithm for the automated extraction of structured data from figures (the orange box on Fig. 1). First, we identified the portion of the image with the axes and legend regions. Second, we looked at axes specifically and parsed the scale for the recalculation of pixel positions to units of measurement. Third, we removed the axes, legend, and gray objects (leaving just the curves themselves) and used a color decomposition algorithm to extract colored curve raw data. This procedure is fully automated if each curve is of a different color.

Axes and legend regions identification. We explored two approaches to detecting axis and legend regions: algorithmic and data-driven. Regarding the algorithmic methods, we tried Canny edge detection[89] combined with the Probabilistic Hough line transform[90] or polygon approximation[91]. We successfully found axes lines for 95% of figures in the dataset. However, these traditional methods expected fixed rules for each detected data type, making the algorithm brittle (see SI.2).

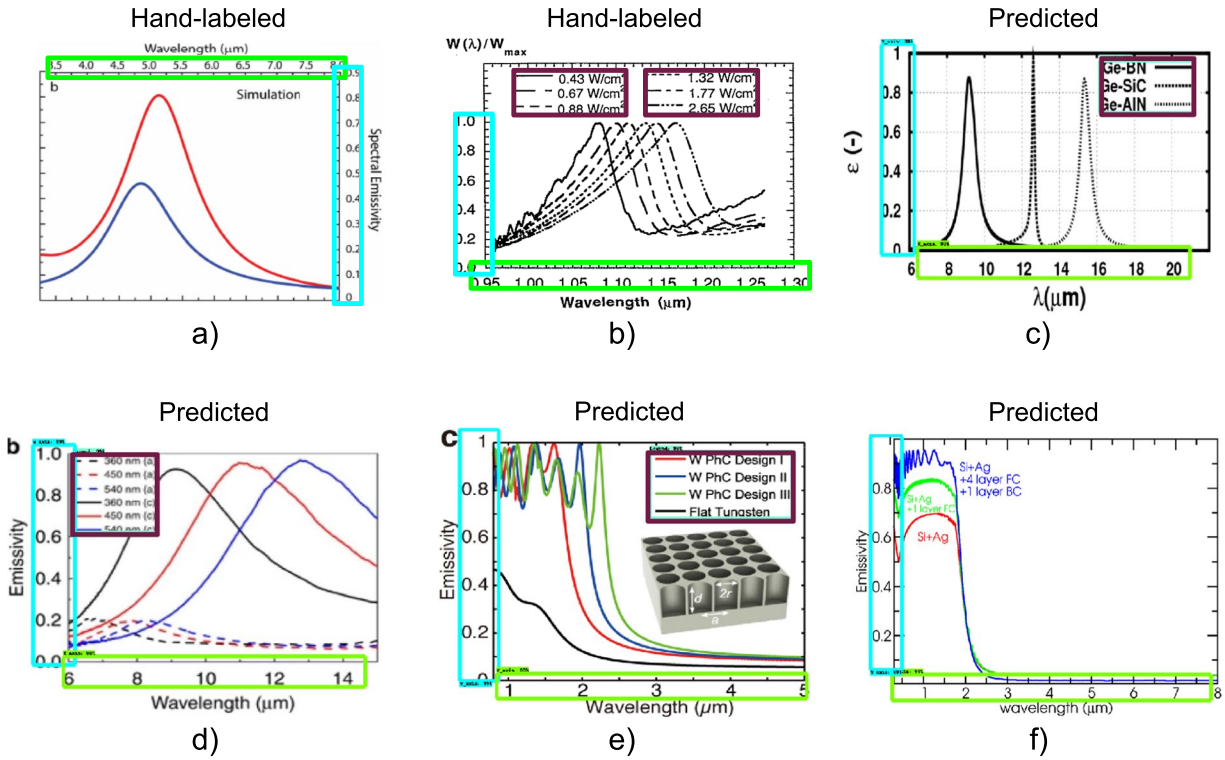

Fig. 3 Examples of axes and legend labeling and trained CNN model performance. The x-axis is outlined in light green, the y-axis in cyan, and the legend in dark magenta. (**a,b**) Examples of hand-labeling using LabelImg[93] software. Boxes depict the identified regions. Note that in b, the y-axis label includes just the portion with numbers and not the entire axis line for subsequent axis scale extraction; see text for details. (**c**–**f**) Examples of output of trained object detection model. Boxes demonstrate the detection results. For a: Copyright 1999–2021 John Wiley and Sons, Inc. All rights reserved. For b: reprinted from Timans, P. J. (1992). The experimental determination of the temperature dependence of the total emissivity of GaAs using a new temperature measurement technique. Journal of applied physics, 72(2), 660–670, with the permission of AIP Publishing. For c: Reprinted from Nefzaoui, E., Drevillon, J., and Joulain, K. (2012). Selective emitters design and optimization for thermophotovoltaic applications. Journal of Applied Physics, 111(8), 084316, with the permission of AIP Publishing.

For the data-driven approach, we used convolutional neural networks[92] (CNNs). CNNs can detect multiple figure features using the same underlying framework but different data labels. We followed a standard procedure for supervised CNN learning with a pre-trained object detection model (first section inside the orange box in Fig. 1). We began by scaling all figures to the size of 800 × 600 pixels (only for the training of CNN; for future steps, we restored the original aspect ratio), splitting the set of 176 figures as 80/20 for training and testing and labeling all figures with LabelImg[93] software. We labeled portions of images corresponding to the three classes: "X_axis", "Y_axis", and "Legend". Axes regions required an axis line, ticks, and numbering. Legend regions included line samples and labels. Figure 3 shows examples of the labelling under a)[21] and b)[80]. In the case of a), we located an X-axis region on top, a Y-axis region to the right, and no legend. Thanks to such images, the trained model can detect the axes objects with numbering on both sides of the axis line. In case b) of the Fig. 3, we identified axes regions containing numbering and two side-by-side legend regions aligned vertically for consistency. Overall, the trained model allows any number of legend objects, including zero.

After compiling the training data, we trained a machine learning model using Tensorflow 2[94] (TF) object detection API[95]. We have a small dataset (for the comparison, the Microsoft COCO 2017[96] object detection dataset has 121408 images), and using a pre-trained model from the Tensorflow Model Zoo is a powerful approach to handle this issue. Among the provided solutions for object detection, we selected faster_rcnn_inception_v2_coco model[97] as it is lightweight with competitive accuracy. The model employs the Faster R-CNN[98] attention mechanism and Inception Resnet[99] deep convolutional network architecture, providing high-speed training. The model was pre-trained on Microsoft COCO 2017 images scaled to 600 × 1024 resolution. Training on our data with the default hyperparameters took 3.5 hours for 10,180 steps on 2.8 GHz Quad-Core Intel Core i7 processor running on a 2019 MacBook Pro.

We evaluated the model performance with standard metrics: precision, recall, and loss based on the intersection-over-union (IOU) method. In object detection tasks, IOU calculates a pixel-by-pixel difference of detected regions from human labels. Then, if we compare the detected region with the corresponding human label and calculate the area of the exact overlap, this value divided by the area of the detected object would

be precision, and divided by the area of the labeled object would be recall. Loss sums up the localization loss (the undetected portion of the label area) and the classification loss (distinguishing between various classes). Following the training, Tensorboard calculated all these metrics on the test set. The model reached 0.75 average precision, 0.81 average recall, and 0.28 average loss (see the description of detection evaluation metrics used by COCO[100] for more details regarding averaging). We note that perfect accuracy on these metrics is not required for the overall task of figure data extraction. Rather, we only need to detect enough of the axis information to be able to correctly perform axis scale parsing (see next step). Manual examination showed that all detected axis objects except one (99.4%) were acceptable in this regard.

Some example results of the model performance are depicted in Fig. 3c–f. In the case of c)[83], the detected x-axis region missed the very left number; nevertheless, the captured information is enough to compute the axis scale, as will be described later. The model detected the y-axis region without issues. We note that the presence of other straight lines like the plot grid did not prevent the algorithm from identifying the axes correctly. Furthermore, the model correctly located the legend despite the unconventional location of the line samples to the right from the labels. In the case of d)[47], all classes were correctly detected: x and y-axis regions contained axis lines, ticks, and all numbers, and the legend object included line samples and labels excepting the borderline. Case e)[101] also had all objects of interest accurately located by the model. In f)[68], the model slightly cropped the x-axis region, missing the last digit but capturing the majority of the numbering; the y-axis was fully detected.

Automated axis scale parsing. Following the identification of the axes and legend regions, the second step in the automatic data retrieval from the images (second section inside the orange box on Fig. 1) was obtaining the axis scale for recalculating pixel positions to units of measurement. We found that optical character recognition[102] (OCR) methods were effective at axis numbering recognition with relatively few modifications or training needed. As the basis of our OCR strategy, we selected EasyOCR[103]. EasyOCR uses Pytorch[104] for deep learning portions and Character-Region Awareness For Text[105] for detection. For recognition, EasyOCR uses Convolutional Recurrent Neural Networks[106] based on ResNet[107], Long Short-Term Memory sequence labelling and Connectionist Temporal Classification[108] decoding with the deep text recognition benchmark[109]. We adjusted EasyOCR's model parameters, imposing a minimum height for the characters of 5 pixels and limiting character detection to numeric characters and special symbols such as a minus symbol or period. Also, we added white padding of 10 pixels from all sides of the images to reduce image edge impact and allow convolutions to operate better. Nevertheless, we note that some fonts were particularly unreadable for the model. EasyOCR returned a list of recognized numbers with the number value and pixel coordinates of the box with the number for every axis region. It properly handled 90% of our figures, producing sufficient information for scale calculation (detecting at least three numbers) for both axes.

To complete the axis scale recalculation, we cleaned EasyOCR output from the poorly detected values as follows. From the EasyOCR output, we filtered out entries with empty or non-numeric text and entries with the probability of recognition lower than 50%. Assuming that the numbers were centered correctly inside the detected boxes and on the tick lines (see SI.3), we approximated the location of ticks in the middle of the detected number boxes. Then, with automated rule-based approach, we applied a polynomial fitting for a set of tick pixel coordinates vs. recognized numbers, ensuring the fitting error subsided 5% for each instance and dropping outliers. We picked two middle points from the accepted set as they are typically more accurate than the edge ones and defined a linear equation for converting pixel coordinates to units of measurement.

Automated curve data extraction. For the final step of the automated figure data extraction, we parsed the colored emissivity curves using image color decomposition (third section inside the orange box on Fig. 1). We chose color decomposition because of the sophistication of this approach, although it has a room for improvement: it does not consider black curves, resulting in about half of the curves being excluded. Figure 4 outlines the framework of the color-based decomposition strategy. The first goal is to remove any extraneous plot elements apart from the data curves themselves. We removed black and gray objects such as axes, text, etc., by transforming images to the Hue Saturation Value (HSV)[110] color scheme and whitening pixels with a Value or Saturation less than 50%. Also, we removed legends detected in the previous processing steps (dark magenta box on Fig. 4a)[23] by coloring them white. This resulted in an image isolated to only colored pixels (top snippet on Fig. 4b). Second, we separated the obtained color-isolated images into clusters of different colors using k-means[111] clustering (bottom snippet on Fig. 4b). We noticed that most of the existing solutions, such as the scikit-learn[112] k-means package or Dominant Color Detection[113], missed the colors corresponding to data curves (see SI.4). Therefore, we adjusted the k-means algorithm initialization (see SI.4). When examining the resulting clusters, some of the color clusters contained single curve records (clusters 3 and 5 on Fig. 4c), some had multiple curves information, and the rest had noisy data (cluster 4 on Fig. 4c). To filter some of the noisy data clusters, we removed clusters for which there were multiple y values for a single x value for more than 1/3 of the x data range. For example, this would exclude a color cluster in which both a solid blue curve and a dashed blue curve were present since the multiple curves would be detected as multiple y values for a single x value. Overall, we obtained clean raw data in pixels for 199 distinct curves.

Finally, we used the calculated axis scales (snippets on the sides of Fig. 4a) to convert pixel positions to units of measurement for the extracted curves. This results in 153 curve records in physical units (orange dataset component on Fig. 1), which determines the total dataset size. We note that we counted every pixel participating in a curve as a data point and did not perform any smoothing operations.

Dataset generation. We applied the data extraction methods described above to the corpus to generate a dataset of thermal emissivity curves and associated metadata. In the case of multiple curves in a single image, we

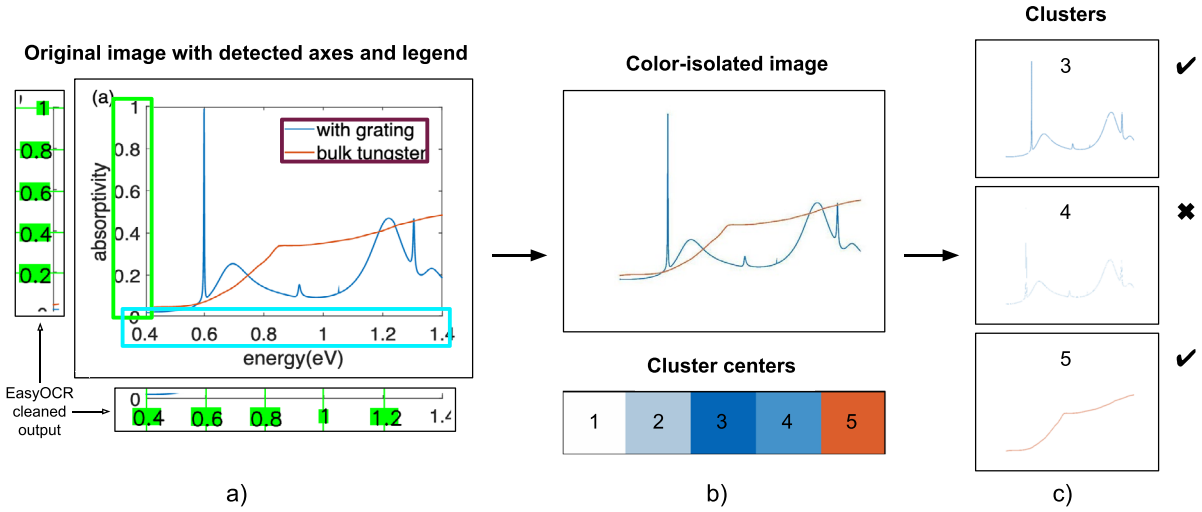

Fig. 4 The pipeline for extracting axis scale and curves of different colors from figures. (**a**) Original image with detected x-axis (light green box), y-axis (cyan box), and legend (dark magenta box). Along the edges of the original image, we show the detected axes regions, the axes scale numbers as detected by EasyOCR[103], and assigned green ticks. (**b**) On top is a color-isolated image that is the original image after removing the axes, legend, and black/gray objects. On the bottom is a color-isolated image palette with cluster centers determined by k-means clustering. (**c**) Data clusters of each color from the palette. Clusters 3 and 5 were accepted as they contain a single curve data. Cluster 4 was rejected as it contained only noise. After extracting the pixel coordinates of clusters 3 and 5, we matched them with EasyOCR cleaned output and converted to units of measurement.

matched curves to metadata manually using a color name reference in the text. We normalized y-coordinates with values from 0 to 1 and standardized x-coordinates to micrometers. The total number of dataset records (one per curve) is 153.

Unsupervised clustering of curve data records. To classify the 153 obtained curves into a manageable number of distinct groups, we applied an unsupervised clustering to the emissivity-wavelength curve data records stored in dataset. We set the wavelength range from 0 μm to 30 μm for all curve data records and reshaped the arrays to a size of 1000. Thus, each curve record became a 1-dimensional array of length 1000 with values from 0 to 1, where zero represented no emissivity measured. Next, we performed unsupervised learning with the DBSCAN algorithm implemented in the Scikit-learn library with parameters eps = 2.6, min_sample = 5 (see SI.5 for the hyperparameter search). The DBSCAN clustered half of the records (all the noise curves were put into a single class for the subsequent analysis, see SI.6) into seven classes of curves with close profiles in terms of Euclidian distance. More details regarding the results are in the Use case section.

Data Records

The dataset of thermal emissivity records with metadata is represented as a set of JSON files and may be found at Figshare[114]. Table 1 provides an overview of the data record schema. The first set of keys refers to article-related attributes: DOI, title, authors, publisher, URL, and year of publication, and also the figure number given in the paper. The remaining attributes are curve-related: list of materials, keywords describing geometry, measurement or calculation method, legend information, axes units, color, score (see Technical Validation for details), and curve raw data. One JSON file corresponds to information retrieved from a single curve. See SI.7 for possible values for various keys.

Technical Validation

We evaluated the efficiency (recall) of the automated figure data extraction pipeline by the portion of the curves extracted from the total number of curves in the data set. Algorithm obtained 153 single curve records. The studied images contained 550 curves, of which half curves were colored curves. The total efficiency over all curves was thus 27%, whereas the total efficiency over colored curves was 55%.

We also studied the quality of the extracted curve data records. To define a quality score, we considered two types of failures in the curve record: gaps in the data and multiple (conflicting) points. Data gaps in a curve frequently occur due to overlapping objects of different colors on the extracted curve and result in x coordinates without corresponding y values. Multiple conflicting points typically appear when the original image had text comments or symbols of the same color as the curve. Such data points contain multiple y values (taking into account the line thickness) for a single x coordinate. We note briefly that attempts to clean text with OCR algorithms often produced gaps; for example, OCR assigned the oscillating portions of data curves with the letter M with very high confidence scores (see SI.8).

	KEY	DESCRIPTION	DATA TYPE
article-related	doi	DOI of the source paper	String
	title	Title of the source paper	String
	authors	Authors of the source paper	List of strings
	publisher	Name of the publishing group	String
	url	Link to the paper	String
	year	Year of paper publication	Integer
	figure_number	Name of figure appearing in the source paper	String
curve-related	materials	All materials used in the sample	List of strings
	geometry	Keywords from the geometry description in paper	List of strings
	composition_key	Assigned keyword: sandwich or single material	String
	geometry_key	Assigned keyword: one of the 7 geometry classes	String
	data_type	Calculation or experiment	String
	tool	Equipment, software or theoretical approach	List of strings
	info_on_image	Additional information appearing on image	String
	axes_units	Units of X and Y axes	Dictionary
	color	HEX color code of the curve	String
	data	Raw curve data in a form of list of [X,Y] coordinates	List of [float, float]
	score	Score of curve from Technical Validation	Float

Table 1. Format of data records in dataset.

Using the failure types defined above, we assigned every x data point of the record as "correct", "gap", or "multiple". We note that small gaps (running for less than 2% of the curve length) were assigned as "dash" to avoid low scores for the dashed style curves. We evaluated the quality of each curve data record, calculating the portion of X data range with correctly extracted points using the Eq. 1:

$$Score = 1 - \frac{N_{GAP} + N_{MULTIPLE}}{N_{GAP} + N_{DASH} + N_{MULTIPLE} + N_{CORRECT}} \quad (1)$$

We also performed hand labeling (using the WebPlotDigitizer[15] software) for curves at different scores and compared the extracted and actual records. The error of the manual extraction was around 2 pixels representing the click accuracy. Figure 5 plots both manual and automated extractions for four cases. The first example is a curve with a score of 1. Automatically extracted points accurately matched manual extraction. The second example is a curve with a score of 0.91. It contains some gaps and a few instances of multiple y points, but most of the curve is extracted correctly. We assigned this record to be of medium quality. The third is a curve with a score of 0.72. Although we correctly handled the dashed style of the curve, this record originally contained a large text comment producing a significant amount of failure data points. This record is considered a poor extraction. Finally, the curve record with the lowest score of 0.31 contained very large gaps and a massive portion of multiple points among the extracted data. Fortunately, the dataset contained only a few records in such a condition.

We grouped the data curve records by the calculated score: good curves with scores exceeding 0.95, medium curves from 0.8 to 0.95, and poor curves for scores below 0.8. Figure 6 depicts the behavior of each group. Approximately half of the records were good, one-third were medium, and one-fifth of the curve records were bad. There were 40 records with scores above 0.99; the worst entry had a score of 0.31. All in all, the proposed automated curve data extraction algorithm produced 122 (80%) good and medium-quality records.

Use case. Next, we examined the data set to understand the overall distribution of the data. This analysis is plotted in Fig. 7. In our dataset[114], the most often used design was a sandwich film. Because this structure is a 1D multilayer stack of thin films, it is easy to model and fabricate. Nevertheless, tuning the composition of the layers, the number of layers, and the thickness of each layer to obtain the desired radiative properties remains a challenge[24]. Another common design in our dataset was a single material slab with a 2D array of cylindrical cavities on the surface. 2D all-metallic emitters are better suitable for high-temperature applications than multilayer structures due to higher chemical and mechanical stability. Also, a 2D grating provides a higher surface-to-volume ratio increasing the emissivity. Grating period, depth, and shape are commonly used to tune the emittance spectrum[48,115]. Analyzing the materials, we found tungsten to be the most popular choice. This is a reasonable observation. Tungsten has the highest melting point among all pure metals and favorable optical properties for selective emitters, such as high emission up to a cutoff wavelength and very little beyond. That makes tungsten a desirable choice for optical samples operating under high temperatures[29]. However, when considering the complete optical device design, the most common configuration was tantalum film with a 2D array of cylindrical cavities on the surface. Tantalum has similar optical properties to tungsten with a high melting point, low vapor pressure, and long-wavelength emissivity (above 2 μm). Also, it is weldable and machinable[48]. Other common configurations include using silica as a sandwich layer[34] and the use of tungsten films with a 2D array of cylindrical cavities[60]. We note, however, that these attributes may change depending on the particular data set of papers used as the source.

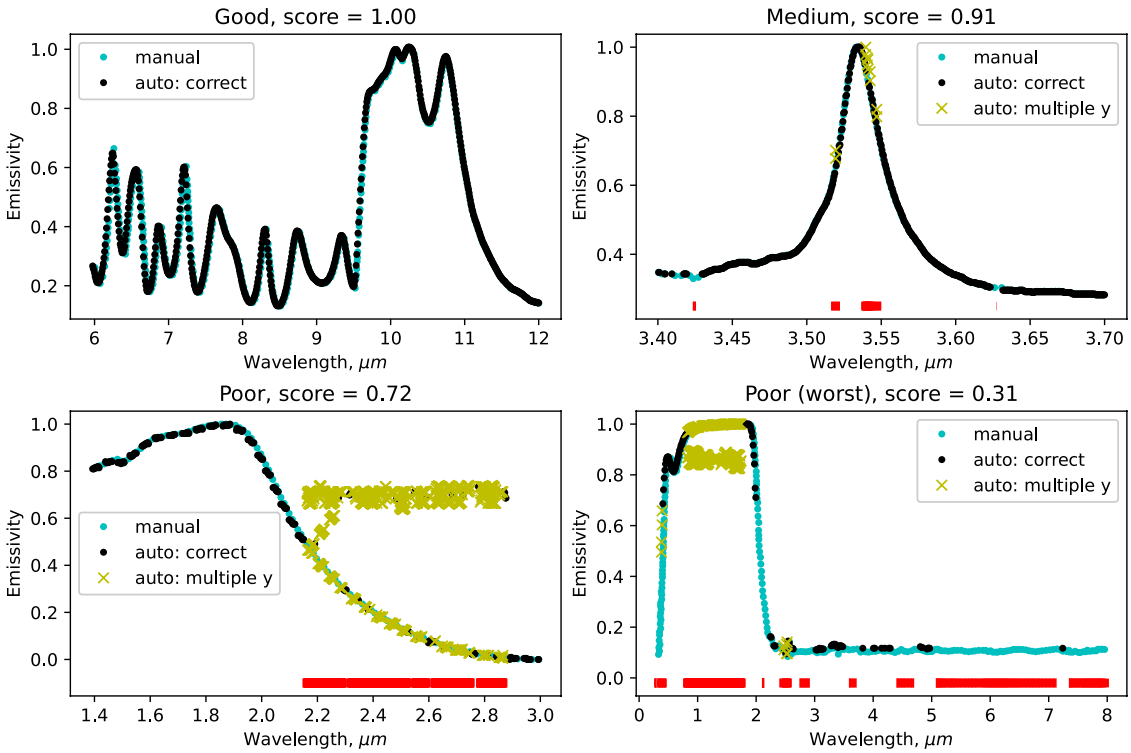

Fig. 5 Several example curves and comparison between automated (black dots for correct points and yellow crosses for multiple y points) and manual (cyan) extraction. The red line on the bottom depicts the unconfident area, demonstrating the portion of the curve where an extraction has failed (gaps, multiple y values). The scores were calculated with Eq. 1. Top left: good extraction, score 1.00, the algorithm correctly captured the entire curve region. Top right: medium quality of extraction, score 0.91, the record has a few gaps and multiple y points. Bottom left: poor extraction, score 0.72, the original curve has dashed style; many multiple y points created by text comment. Bottom right: poor extraction, the lowest score of 0.31; many gaps caused by overlapping, multiple y points due to text comment.

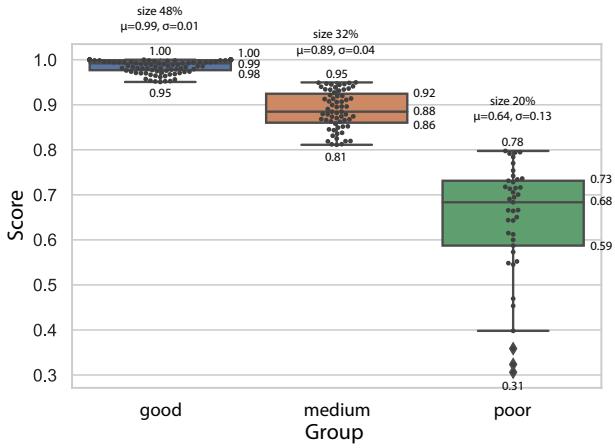

Fig. 6 Statistical analysis of extracted curve data records grouped as good, medium, and poor. Each point represents one curve, and scores show the quality of the extraction. Scores were calculated with the Eq. 1. A mean μ and standard deviation σ values are given on top, along with the relative size of each group.

Next, we aimed to find trends in emissivity profiles and correlate them with device attributes. We followed an unsupervised clustering strategy to identify groups of emissivity curves with similar behavior (see Methods section) and analyzed metadata within each group. Figure 8 shows that the agreement inside each group is generally good: curves are plotted with partial transparency, and the darker regions correspond to curve overlap. Each class depicts a unique emissivity behavior. We note that class 6 had a variety of samples and therefore was nonuniform. All classes except 6 had a single dominant design and composition with few outliers (pie insets in Fig. 8) as well as dominant materials (bar insets in Fig. 8). We observed that class 1 had the sharpest peak compared to

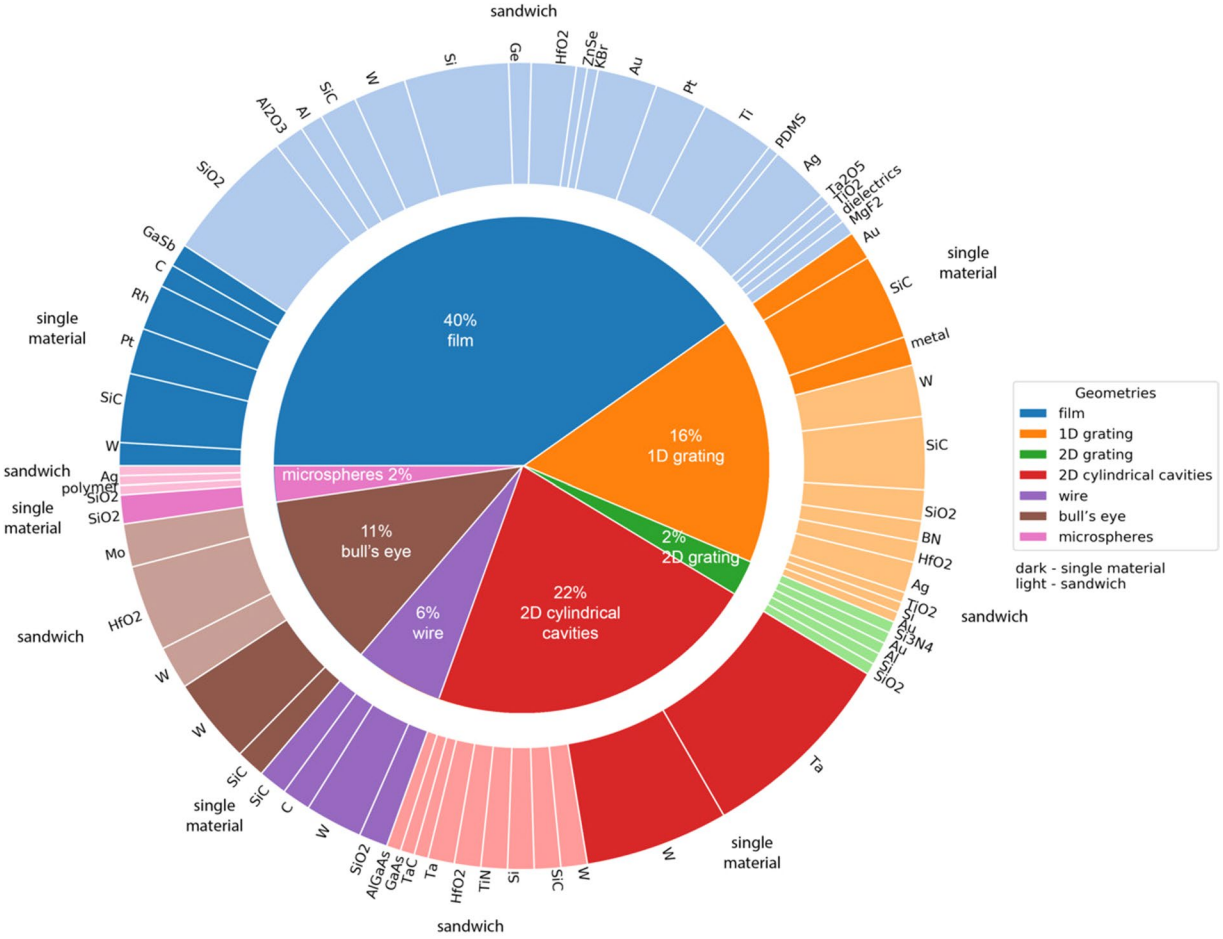

Fig. 7 The distribution of design-related parameters (geometry, materials) in the dataset. The innermost circle corresponds to geometry. The outer ring depicts the used materials with colors reflecting the composition: the color is dark for single material devices and light for sandwich structures. There are 32 distinct materials. In total, there are 60% sandwich and 40% single material structures. The most used material overall is tungsten, which has desirable properties for optical devices.

others. Most class members had a bull's eye structure which is indeed designed for thermal beaming. A series of equally spaced circular concentric grooves produces an emission spectrum in the normal direction with a single peak at a wavelength nearly identical to the period[62]. Class 2 forms a bi-modal emissivity profile. Sandwich films, in this case, were designed as Fabry-Perot cavity resonators[116]. They contained *Si*, *Ti*, and *Pt* layers covered with opaque (thick) *Au* layer, SiO_2 cavity layer, ultra-thin top *Au* layer, and SiO_2 protection layer. Fabry-Perot cavity resonators produce two emission peaks at locations determined by the optical properties of the cavity, opaque and top layer materials (SiO_2, *Au*), and the cavity's thickness. Peaks amplitudes are sensitive to the thickness of the top layer[54].

As Fig. 8 demonstrates, geometry seems to be the major factor defining curve behavior. Designs without any grating on the surface ended up in classes 2 and 3. Records with a bull's eye geometries fell in class 1. If there was a 2D periodic grating on the surface, we obtained classes 4, 5, and 7. The selection of materials further defined the curve behavior. Usage of *Au* and presence of SiO_2 cavity determined class 2, while class 3 members were multilayer stacks of cermet layers with a *Ag* reflective back[29]. Similarly, *Ta* was characteristic for class 4, *W* for class 5, and *TiN* for class 7. We trained a decision tree (see SI.9) that further demonstrated primary splitting of behavior on geometrical attributes and secondary splitting on a choice of materials.

Other work. A dataset of emissivity curves was previously reported by Frolec *et al*. in 2018[117]. It contains 58 records of thermal emissivities experimentally measured by authors starting at cryogenic temperatures and slightly exceeding room temperature. The curves were obtained from 45 different samples covering a range of pure metals, alloys, foils, coated metals, and ceramic plates. The dataset does not contain spectral data but provides information on the bulk material, coating layer material, treatment techniques, and the temperature dependence of the hemispherical emissivity or absorptivity. In contrast to the current work, the information was not compiled from the literature but rather was measured by the authors. Thus, this dataset represents more consistent techniques for data generation but is more limited in scope and delivers less information.

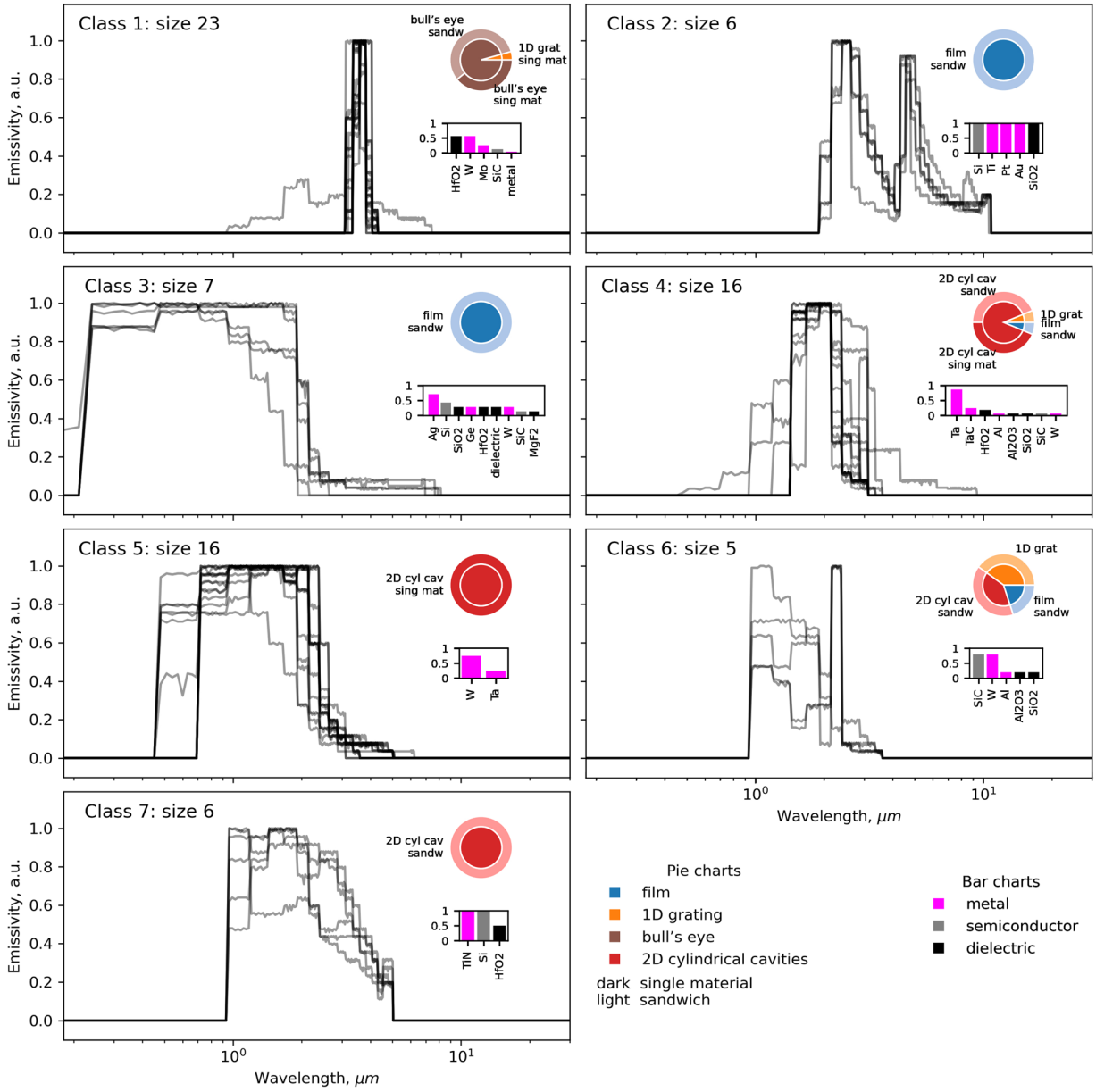

Fig. 8 Curves classes with similar emissivity behavior and distribution of corresponding metadata. Curves were clustered with unsupervised learning using the DBSCAN algorithm. Curves are plotted with partial transparency such that dark areas indicate overlap of curves. The x-axis is in logarithm scale for better visualization. Pie charts in the insets show the distribution of geometry and composition per class. Bar charts in the insets depict distinct material frequencies normalized per class size (i.e., if the bin height is 1, the material is present in every record in the class).

Another work from Kobayashi *et al*. presents normal spectral emissivity dataset measured at high temperatures[118] reaching 1500 K. This work also includes the results of measurements performed by the authors in the same laboratory and the dataset does not provide spectral data. All the investigated materials were metal surfaces with different degrees of oxidation, and the surface roughness is stored as one of the parameters in the records. Seven different metals were studied, and only a single design was represented. Thus, while the consistency of the method is higher than the one we report, the scale and diversity of data are more limited.

Our dataset contains spectral data and corresponding materials, method and design parameters. In our dataset, the temperatures usually lie between room temperature and 2500 K, although we have not rigorously parsed all the temperature values for all curves (information regarding the temperature is sometimes contained within the "info_on_image" key in the JSON records). Emissivity - wavelength data relations are both experimental and theoretical obtained with different equipment. We store many designs of different complexity and details regarding the sample geometry. There are 32 distinct materials. Thus, our dataset does not have any records fully duplicating those mentioned above or any other work, advancing previously published emissivity databases.

In this work, we manually collected 64 papers with 176 figures reporting emissivity and automatically extracted 153 curve records with manual extraction of corresponding metadata. However, there exist more publications reporting emissivity. We analyzed the collected metadata in our dataset with the CountVectorizer package featured in Scikit-learn library without tokenization[119] and found that only 45% of captions of relevant figures mention emissivity, while 88% of captions mention any of emissivity, emission or emittance (9% mention both emissivity and emission, 17% mention only emission, 26% mention only emittance). Thus, among relevant entries, 88% satisfy the criteria of three words; and half of interesting figures may be missed if we excluded terms emission or emittance from the search. Then, we checked manually that among 100 random figures satisfying the criteria of three words (mentioning emissivity, emission or emittance in the figure caption), 70% indeed presented thermal emissivity curves. To determine how many more figures might be possible to collect, we implemented nine electronic paper scrapers (see SI.10) that checked figure captions amongst 4.9 million papers for the above keywords. As an aside, we note that there are packages like EXSCLAIM![120] to help automate this process, but it does not support the journal publishers targeted in our work. Our results indicate that there exist 361,000 figures (178,000 papers) mentioning emissivity, emission, or emitter in the caption and potentially 70% of them would have the data we want. While we envision it may be possible to use an automatic curve data extraction algorithm to obtain a reasonable fraction of these curves, extracting the design-related parameters from text (i.e., the geometries, materials, and methods that describe each curve) remains a challenge and the biggest bottleneck for expanding the optical properties databases.

The automated generation of databases incorporating textual and spectral data has several remaining challenges. First is corpus composition, which includes an automated selection of papers relevant to specific tasks; it can be accommodated with existing text-mining tools[121]. The second is metadata extraction, and advanced text-mining algorithms fine-tuned for specific applications are promising for this task[122]. The next challenge comes from curve extraction with a color decomposition strategy; as described previously, OCR routines fail to distinguish between curve and text comments of the same color and, therefore, cannot be used to exclude text data. Furthermore, black curves cannot be isolated since the curve detection routine removes all grayscale pixels (e.g., to remove axes) prior to curve isolation. Also, dashed and solid curves of the same color cannot be differentiated. More advanced methods, such as image segmentation, may be able to overcome some of these limitations.[94]. The other is extracting the additional information from figures, such as linking legend labels with curves, that has been partially addressed in other methods[20].

Usage Notes

The set of JSON files is available at Figshare[114]. Each file can be opened with any software for text editing or by common programming languages. The python script for re-plotting the data from any of the JSON records is available at https://github.com/ViktoriiaBaib/curvedataextraction and called "replot_DBrecord.py". The repository also contains some scripts for querying the dataset for the presented analysis and beyond.

Code availability

The source code (implemented in Python) for performing all the described figure analysis steps and generating the data entries is available at https://github.com/ViktoriiaBaib/curvedataextraction. The axis and legend detection step uses the TensorFlow2 Object Detection API and provides a fine-tuned CNN model. File "object_detection_axes_legend.py" performs object detection of legend, x-axis, and y-axis objects and generates PNG and JSON records for these objects. File "color_decomposition.py" performs clustering by color and produces PNG of color-isolated image, palette, as well as PNG and JSON records of separate color clusters in pixel coordinates. It uses methods from "posterization.py". File "final-record.py" performs axes scale parsing and applies it to all the clusters, producing cluster records in units of measurement. It utilizes methods from "final_record_func.py".

Received: 10 June 2022; Accepted: 14 September 2022;
Published online: 29 September 2022

References

1. Fritts, C. E. On a new form of selenium cell, and some electrical discoveries made by its use. *American Journal of Science* **s3-26**, 465–472, https://doi.org/10.2475/ajs.s3-26.156.465 (1883).
2. Solomon, M. L. *et al*. Nanophotonic platforms for chiral sensing and separation. *Accounts of chemical research* **53**, 588–598 (2020).
3. Ito, T. & Okazaki, S. Pushing the limits of lithography. *Nature* **406**, 1027–1031 (2000).
4. Krebs, H.-U. *et al*. Pulsed laser deposition (pld)–a versatile thin film technique. In *Advances in Solid State Physics*, 505–518 (Springer, 2003).
5. Kunz, K. S. & Luebbers, R. J. *The finite difference time domain method for electromagnetics* (CRC press, 1993).
6. Ling, H., Li, R. & Davoyan, A. R. All van der waals integrated nanophotonics with bulk transition metal dichalcogenides. *ACS Photonics* **8**, 721–730 (2021).
7. Blokhin, E. & Villars, P. The pauling file project and materials platform for data science: From big data toward materials genome. *Handbook of Materials Modeling: Methods: Theory and Modeling* 1837–1861 (2020).
8. Rothman, L. S. History of the hitran database. *Nature Reviews Physics* **3**, 302–304 (2021).
9. Kim, E. *et al*. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chemistry of Materials* **29**, 9436–9444 (2017).
10. Kononova, O. *et al*. Text-mined dataset of inorganic materials synthesis recipes. *Scientific data* **6**, 1–11 (2019).
11. Huang, S. & Cole, J. M. A database of battery materials auto-generated using chemdataextractor. *Scientific Data* **7**, 1–13 (2020).
12. Dong, Q. & Cole, J. M. Auto-generated database of semiconductor band gaps using chemdataextractor. *Scientific Data* **9**, 1–11 (2022).
13. Zhao, J. & Cole, J. M. A database of refractive indices and dielectric constants auto-generated using chemdataextractor. *Scientific data* **9**, 1–11 (2022).
14. Katsura, Y. *et al*. Data-driven analysis of electron relaxation times in pbte-type thermoelectric materials. *Science and Technology of Advanced Materials* **20**, 511–520 (2019).

15. Rohatgi, A. Webplotdigitizer: Version 4.4. https://automeris.io/WebPlotDigitizer/ (2020).
16. GRABIT, MATLAB Central File Exchange. https://www.mathworks.com/matlabcentral/fileexchange/7173-grabit/ (2021).
17. Tummers, B. DataThief III. https://datathief.org/ (2006).
18. Jiang, W. *et al*. Plot2spectra: an automatic spectra extraction tool. *arXiv preprint arXiv:2107.02827* (2021).
19. Clark, C. A. & Divvala, S. Looking beyond text: Extracting figures, tables and captions from computer science papers. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence* (2015).
20. Siegel, N., Horvitz, Z., Levin, R., Divvala, S. & Farhadi, A. Figureseer: Parsing result-figures in research papers. In Leibe, B., Matas, J., Sebe, N. & Welling, M. (eds.) *Computer Vision–ECCV 2016*, 664–680 (Springer International Publishing, Cham, 2016).
21. Liu, X. & Padilla, W. J. Thermochromic infrared metamaterials. *Advanced Materials* **28**, 871–875 (2016).
22. Narayanaswamy, A., Cybulski, J. & Chen, G. 1d metallo-dielectric photonic crystals as selective emitters for thermophotovoltaic applications. In *AIP Conference Proceedings*, vol. 738, 215–220 (American Institute of Physics, 2004).
23. Guo, Y. & Fan, S. Narrowband thermal emission from a uniform tungsten surface critically coupled with a photonic crystal guided resonance. *Optics express* **24**, 29896–29907 (2016).
24. Zeyghami, M., Stefanakos, E. & Goswami, D. Y. Development of one-dimensional photonic selective emitters for energy harvesting applications. *Solar Energy Materials and Solar Cells* **163**, 191–199 (2017).
25. Rephaeli, E. & Fan, S. Absorber and emitter for solar thermo-photovoltaic systems to achieve efficiency exceeding the shockley-queisser limit. *Optics express* **17**, 15145–15159 (2009).
26. Sai, H., Kanamori, Y. & Yugami, H. High-temperature resistive surface grating for spectral control of thermal radiation. *Applied Physics Letters* **82**, 1685–1687 (2003).
27. Timans, P. Emissivity of silicon at elevated temperatures. *Journal of Applied Physics* **74**, 6353–6364 (1993).
28. Thomas, N. H., Chen, Z., Fan, S. & Minnich, A. J. Semiconductor-based multilayer selective solar absorber for unconcentrated solar thermal energy conversion. *Scientific reports* **7**, 1–6 (2017).
29. Chester, D., Bermel, P., Joannopoulos, J. D., Soljacic, M. & Celanovic, I. Design and global optimization of high-efficiency solar thermal systems with tungsten cermets. *Optics express* **19**, A245–A257 (2011).
30. Wang, H., Kaur, S., Elzouka, M. & Prasher, R. A nano-photonic filter for near infrared radiative heater. *Applied Thermal Engineering* **153**, 221–224 (2019).
31. Golyk, V. A., Krüger, M. & Kardar, M. Heat radiation from long cylindrical objects. *Physical Review E* **85**, 046603 (2012).
32. Costantini, D. *et al*. Plasmonic metasurface for directional and frequency-selective thermal emission. *Physical Review Applied* **4**, 014023 (2015).
33. Ghebrebrhan, M. *et al*. Tailoring thermal emission via q matching of photonic crystal resonances. *Physical Review A* **83**, 033810 (2011).
34. Sakurai, A. *et al*. Ultranarrow-band wavelength-selective thermal emission with aperiodic multilayered metamaterials designed by bayesian optimization. *ACS central science* **5**, 319–326 (2019).
35. King, J. L. *et al*. Impact of corrosion on the emissivity of advanced reactor structural alloys. *Journal of Nuclear Materials* **508**, 465–471 (2018).
36. Sergeant, N. P., Pincon, O., Agrawal, M. & Peumans, P. Design of wide-angle solar-selective absorbers using aperiodic metal-dielectric stacks. *Optics express* **17**, 22800–22812 (2009).
37. Sani, E., Mercatelli, L., Fontani, D., Sans, J.-L. & Sciti, D. Hafnium and tantalum carbides for high temperature solar receivers. *Journal of Renewable and Sustainable Energy* **3**, 063107 (2011).
38. Hervé, A., Drévillon, J., Ezzahri, Y. & Joulain, K. Radiative cooling by tailoring surfaces with microstructures: Association of a grating and a multi-layer structure. *Journal of Quantitative Spectroscopy and Radiative Transfer* **221**, 155–163 (2018).
39. He, X., Li, Y., Wang, L., Sun, Y. & Zhang, S. High emissivity coatings for high temperature application: Progress and prospect. *Thin Solid Films* **517**, 5120–5129 (2009).
40. Leroy, A. *et al*. High performance incandescent lighting using a selective emitter and nanophotonic filters. In *Thermal Radiation Management for Energy Applications*, vol. 10369, 41–51 (SPIE, 2017).
41. Nefzaoui, E., Drevillon, J. & Joulain, K. Nanostructures thermal emission optimization using genetic algorithms and particle swarms. In *International Conference on Evolutionary Computation 2010 (ICEC 2010)*, 219–224 (2010).
42. Zhai, Y. *et al*. Scalable-manufactured randomized glass-polymer hybrid metamaterial for daytime radiative cooling. *Science* **355**, 1062–1066 (2017).
43. Cagran, C. P., Hanssen, L. M., Noorma, M., Gura, A. V. & Mekhontsev, S. N. Temperature-resolved infrared spectral emissivity of sic and pt–10rh for temperatures up to 900°c. *International Journal of Thermophysics* **28**, 581–597 (2007).
44. Argyropoulos, C., Le, K. Q., Mattiucci, N., D'Aguanno, G. & Alu, A. Broadband absorbers and selective emitters based on plasmonic brewster metasurfaces. *Physical Review B* **87**, 205112 (2013).
45. Trotter, D. Jr & Sievers, A. Thermal emissivity of selective surfacesâ€"new lower limits. *Applied Physics Letters* **35**, 374–376 (1979).
46. Ben-Abdallah, P. & Ni, B. Single-defect bragg stacks for high-power narrow-band thermal emission. *Journal of applied physics* **97**, 104910 (2005).
47. Du, K.-K. *et al*. Control over emissivity of zero-static-power thermal emitters based on phase-changing material gst. *Light: Science & Applications* **6**, e16194–e16194 (2017).
48. Rinnerbauer, V. *et al*. Large-area fabrication of high aspect ratio tantalum photonic crystals for high-temperature selective emitters. *Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena* **31**, 011802 (2013).
49. Nefedov, I. S. & Melnikov, L. A. Super-planckian far-zone thermal emission from asymmetric hyperbolic metamaterials. *Applied Physics Letters* **105**, 161902 (2014).
50. Chan, W. R. *et al*. Toward high-energy-density, high-efficiency, and moderate-temperature chip-scale thermophotovoltaics. *Proceedings of the National Academy of Sciences* **110**, 5309–5314 (2013).
51. Sakr, E. & Bermel, P. Angle-selective reflective filters for exclusion of background thermal emission. *Physical Review Applied* **7**, 044020 (2017).
52. Sai, H. & Yugami, H. Thermophotovoltaic generation with selective radiators based on tungsten surface gratings. *Applied physics letters* **85**, 3399–3401 (2004).
53. Lee, B., Fu, C. & Zhang, Z. Coherent thermal emission from one-dimensional photonic crystals. *Applied Physics Letters* **87**, 071904 (2005).
54. Wang, L., Basu, S. & Zhang, Z. Direct measurement of thermal emission from a fabry–perot cavity resonator. *Journal of heat transfer* **134** (2012).
55. Rinnerbauer, V. *et al*. Superlattice photonic crystal as broadband solar absorber for high temperature operation. *Optics express* **22**, A1895–A1906 (2014).
56. Kou, J.-l, Jurado, Z., Chen, Z., Fan, S. & Minnich, A. J. Daytime radiative cooling using near-black infrared emitters. *Acs Photonics* **4**, 626–630 (2017).
57. Lee, H.-J. *et al*. Hafnia-plugged microcavities for thermal stability of selective emitters. *Applied Physics Letters* **102**, 241904 (2013).
58. Lenert, A. *et al*. 2d photonic-crystals for high spectral conversion efficiency in solar thermophotovoltaics. In *2014 IEEE 27th International Conference on Micro Electro Mechanical Systems (MEMS)*, 576–579 (IEEE, 2014).
59. Wang, L. & Zhang, Z. Phonon-mediated magnetic polaritons in the infrared region. *Optics express* **19**, A126–A135 (2011).

60. Li, W. & Fan, S. Nanophotonic control of thermal radiation for energy applications. *Optics express* **26**, 15995–16021 (2018).
61. Rostamnejadi, A. & Daneshvar, M. Two-dimensional tungsten photonic crystal selective emitter: effects of geometrical parameters and temperature. *Applied Physics B* **124**, 1–8 (2018).
62. Park, J. H., Han, S. E., Nagpal, P. & Norris, D. J. Observation of thermal beaming from tungsten and molybdenum bull's eyes. *ACS Photonics* **3**, 494–500 (2016).
63. Mandal, J. *et al*. Hierarchically porous polymer coatings for highly efficient passive daytime radiative cooling. *Science* **362**, 315–319 (2018).
64. Cao, F., McEnaney, K., Chen, G. & Ren, Z. A review of cermet-based spectrally selective solar absorbers. *Energy & Environmental Science* **7**, 1615–1627 (2014).
65. Rinnerbauer, V. *et al*. High-temperature stability and selective thermal emission of polycrystalline tantalum photonic crystals. *Optics express* **21**, 11482–11491 (2013).
66. Zhu, L., Raman, A., Wang, K. X., Abou Anoma, M. & Fan, S. Radiative cooling of solar cells. *Optica* **1**, 32–38 (2014).
67. Ilic, O. *et al*. Tailoring high-temperature radiation and the resurrection of the incandescent source. *Nature nanotechnology* **11**, 320–324 (2016).
68. Bermel, P. *et al*. Design and global optimization of high-efficiency thermophotovoltaic systems. *Optics express* **18**, A314–A334 (2010).
69. Celanovic, I., Perreault, D. & Kassakian, J. Resonant-cavity enhanced thermal emission. *Physical Review B* **72**, 075127 (2005).
70. Arpin, K. A. *et al*. Three-dimensional self-assembled photonic crystals with high temperature stability for thermal emission modification. *Nature communications* **4**, 1–8 (2013).
71. Biener, G., Dahan, N., Niv, A., Kleiner, V. & Hasman, E. Highly coherent thermal emission obtained by plasmonic bandgap structures. *Applied Physics Letters* **92**, 081913 (2008).
72. Yeng, Y. X. *et al*. Performance analysis of experimentally viable photonic crystal enhanced thermophotovoltaic systems. *Optics express* **21**, A1035–A1051 (2013).
73. Baranov, D. G. *et al*. Nanophotonic engineering of far-field thermal emitters. *Nature materials* **18**, 920–930 (2019).
74. Busch, K. *et al*. Periodic nanostructures for photonics. *Physics reports* **444**, 101–202 (2007).
75. Shi, Y., Li, W., Raman, A. & Fan, S. Optimization of multilayer optical films with a memetic algorithm and mixed integer programming. *Acs Photonics* **5**, 684–691 (2017).
76. Chan, D. L., Soljačić, M. & Joannopoulos, J. Thermal emission and design in one-dimensional periodic metallic photonic crystal slabs. *Physical Review E* **74**, 016609 (2006).
77. De Zoysa, M. *et al*. Conversion of broadband to narrowband thermal emission through energy recycling. *Nature Photonics* **6**, 535–539 (2012).
78. Raman, A. P., Anoma, M. A., Zhu, L., Rephaeli, E. & Fan, S. Passive radiative cooling below ambient air temperature under direct sunlight. *Nature* **515**, 540–544 (2014).
79. DeSutter, J., Bernardi, M. P. & Francoeur, M. Determination of thermal emission spectra maximizing thermophotovoltaic performance using a genetic algorithm. *Energy Conversion and Management* **108**, 429–438 (2016).
80. Timans, P. The experimental determination of the temperature dependence of the total emissivity of gaas using a new temperature measurement technique. *Journal of applied physics* **72**, 660–670 (1992).
81. Atiganyanun, S. *et al*. Effective radiative cooling by paint-format microsphere-based photonic random media. *ACS Photonics* **5**, 1181–1187 (2018).
82. Rinnerbauer, V. *et al*. Recent developments in high-temperature photonic crystals for energy conversion. *Energy & Environmental Science* **5**, 8815–8823 (2012).
83. Nefzaoui, E., Drevillon, J. & Joulain, K. Selective emitters design and optimization for thermophotovoltaic applications. *Journal of Applied Physics* **111**, 084316 (2012).
84. Yeng, Y. X. *et al*. Enabling high-temperature nanophotonics for energy applications. *Proceedings of the National Academy of Sciences* **109**, 2280–2285 (2012).
85. Elsevier. Mendeley ltd. https://www.mendeley.com/ (2021).
86. project, L. The latex project. https://www.latex-project.org/ (2022).
87. AB, S. L. Secret labs' regular expression engine. https://docs.python.org/3/library/re.html/ (2021).
88. Golovizin, A. A bibtex-compatible bibliography processor in python. https://pypi.org/project/pybtex/ (2021).
89. Canny, J. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence* 679–698 (1986).
90. Matas, J., Galambos, C. & Kittler, J. Robust detection of lines using the progressive probabilistic hough transform. *Computer vision and image understanding* **78**, 119–137 (2000).
91. Bradski, G. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer* **25**, 120–123 (2000).
92. Valueva, M. V., Nagornov, N., Lyakhov, P. A., Valuev, G. V. & Chervyakov, N. I. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. *Mathematics and Computers in Simulation* **177**, 232–243 (2020).
93. Tzutalin. LabelImg. https://github.com/tzutalin/labelImg/ (2015).
94. Abadi, M. *et al*. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org (2015).
95. Huang, J. *et al*. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7310–7311 (2017).
96. Lin, T.-Y. *et al*. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755 (Springer, 2014).
97. Grus, J. Tensorflow 1 detection model zoo. https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf1_detection_zoo.md/ (2018).
98. Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **28**, 91–99 (2015).
99. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence* (2017).
100. Lin, T.-Y. Detection evaluation metrics used by coco. https://cocodataset.org/ (2019).
101. Li, W. & Fan, S. Nanophotonic control of thermal radiation for energy applications. *Optics express* **26**, 15995–16021 (2018).
102. Herbert, H. The history of ocr, optical character recognition. *Manchester Center, VT: Recognition Technologies Users Association* (1982).
103. EasyOCR. https://github.com/jaidedai/easyocr/ (2020).
104. PyTorch. https://pytorch.org/ (2020).
105. Baek, Y., Lee, B., Han, D., Yun, S. & Lee, H. Character region awareness for text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9365–9374 (2019).
106. Shi, B., Bai, X. & Yao, C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence* **39**, 2298–2304 (2016).
107. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

108. Graves, A., Fernández, S., Gomez, F. & Schmidhuber, J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, 369–376 (2006).
109. Baek, J. *et al*. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *International Conference on Computer Vision (ICCV)* (2019).
110. Cheng, H.-D., Jiang, X. H., Sun, Y. & Wang, J. Color image segmentation: advances and prospects. *Pattern recognition* **34**, 2259–2281 (2001).
111. Likas, A., Vlassis, N. & Verbeek, J. J. The global k-means clustering algorithm. *Pattern recognition* **36**, 451–461 (2003).
112. Pedregosa, F. *et al*. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
113. Havidek, H. Dominant color detection. https://pypi.org/project/dominant-color-detection/ (2020).
114. Baibakova, V. Optical emissivity dataset of multi-material heterogeneous designs generated with automated figure extraction, *Figshare*, https://doi.org/10.6084/m9.figshare.c.6037004.v1 (2022).
115. Schlemmer, C., Aschaber, J., Boerner, V. & Luther, J. Thermal stability of micro-structured selective tungsten emitters. In *AIP Conference Proceedings*, vol. 653, 164–173 (American Institute of Physics, 2003).
116. Schubert, E. *et al*. Enhanced photoluminescence by resonant absorption in er-doped sio2/si microcavities. *Applied physics letters* **63**, 2603–2605 (1993).
117. Frolec, J. *et al*. A database of metallic materials emissivities and absorptivities for cryogenics. *Cryogenics* **97**, 85–99 (2019).
118. Kobayashi, M., Ono, A., Otsuki, M., Sakate, H. & Sakuma, F. A database of normal spectral emissivities of metals at high temperatures. *International journal of thermophysics* **20**, 299–308 (1999).
119. Grefenstette, G. Tokenization. In *Syntactic Wordclass Tagging*, 117–133 (Springer, 1999).
120. Schwenker, E. *et al*. Exsclaim!–an automated pipeline for the construction of labeled materials imaging datasets from literature. *arXiv preprint arXiv:2103.10631* (2021).
121. Weston, L. *et al*. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of chemical information and modeling* **59**, 3692–3702 (2019).
122. Brown, T. *et al*. Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020).

Acknowledgements

This research was supported by the US Department of Energy Advanced Research Projects Agency-Energy under Contract No. 19/CJ000/04/01. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.

Author contributions

V.B. performed coding, extraction, and analysis work, M.E. collected the initial dataset of relevant papers, A.J. guided the algorithm development and project planning, S.L. contributed to the use case section, and R.P. provided the overall project vision. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-022-01699-3.

Correspondence and requests for materials should be addressed to A.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.