Supplementary information for: Real-time hyperspectral imaging in hardware via trained metasurface encoders

M. Makarenko†, A. Burguete-Lopez, Q. Wang, F. Getman, Silvio Giancola,
Bernard Ghanem & A. Fratalocchi
King Abdullah University of Science and Technology (KAUST)
Thuwal, 23955-6900, KSA
†maksim.makarenko@kaust.edu.sa

1. Details on time-domain coupled-mode theory TDCMT

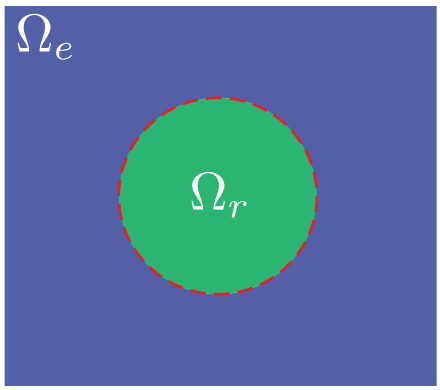

Figure 1. Splitting of metasurface geometry space Ω into resonance space Ω_r and exterior (propagation) space Ω_e

In this section, we describe a set of exact coupled-mode equations that are fully equivalent to Maxwell's equations. The main idea of the coupled-mode approach is to divide the geometrical space Ω where light propagation into a resonator space Ω_r and an external space Ω_e (Fig. 1). We assume that the external space does not contain sources or charges. Under this formulation, the set of Maxwell equations reduces to the following set of exact coupled-mode equations [2]:

$$\begin{cases} \tilde{\mathbf{a}}(\omega) = \frac{\tilde{K}}{i(\omega - W) + \frac{\tilde{K}\tilde{K}^\dagger}{2}} \tilde{\mathbf{s}}_+ \\ \tilde{\mathbf{s}}_-(\omega) = \tilde{C}\left(\tilde{\mathbf{s}}_+ - \tilde{K}^\dagger \cdot \tilde{\mathbf{a}}\right) \end{cases} \tag{1}$$

with $1/\tilde{X}$ the inverse matrix $\tilde{X}^{-1}$. Power conservation implies that the matrix σ :

$$\sigma = \mathbf{1} - \tilde{K} \frac{1}{i(\omega - W) + \frac{\tilde{K}\tilde{K}^\dagger}{2}} \tilde{K}^\dagger \tag{2}$$

defined from the solution of the coupled mode equations, is unitary $\sigma^\dagger \cdot \sigma = 1$.

Equations (1) show that the dynamics of the system depend only on three independent matrices: the coupling matrix $\tilde{K}$, the scattering matrix $\tilde{C}$, and the resonance matrix W.

2. Network training for supervised spectral prediction

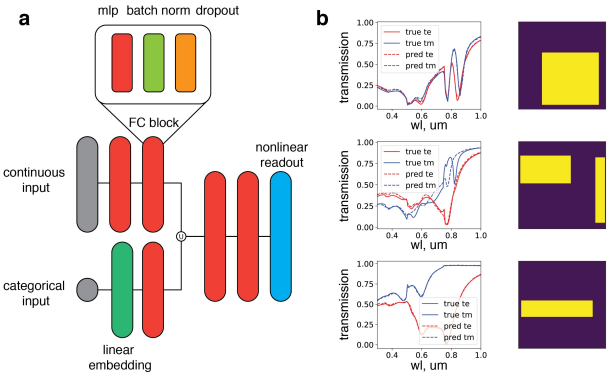

Figure 2. Differentiable ALFRED spectral predictor. **a** Conceptual sketch of d-alfred neural network shape-to-resonance mapper. **b** Qualitative results of spectra prediction for the dataset samples.

Figure 2 illustrates the model of the proposed differentiable spectral predictor (d-ALFRED). It consists of several fully connected (FC) blocks connected sequentially. Each FC block consists of multi-layer perceptrons (MLP) of different sizes, batch normalization layer, and dropout. To process separately categorical variables in input (period, thickness), we design an alternative branch consisting of a linear embedding layer and FC block connected sequentially. The primary purpose of this branch is to balance categorical and continuous variable's weights in the model. Then we concatenate both continuous and categorical features and feed

them into readout blocks, consisting of multiple FC blocks.

We use the training dataset provided by [3], which contains over 600 000 simulation results of pure silicon structures on top of glass under a Total-Field Scattered-Field (TFSF) simulation. Each simulation has periodic boundary conditions with one of the three different periods (250 nm, 500 nm or 750 nm) and one of the ten different discrete thicknesses from 50 to 300 nm with a 25 nm step. Each structure consists of a random combination (up to 5) of cuboid resonators. We split the dataset into test and training parts comprising 20% and 80% of the total, respectively, then take 10% of the training set as a validation set.

For the training part, we use the Adam optimizer [1] with a learning rate 1×10^{-5} and a step learning rate scheduler with $stepsize = 50$ and $\gamma = 0.1$ hyperparameters. For the desired system response in either transmission or reflection, we apply a sigmoid activation function at the top layer of FCN. This function maps the output spectrum to the range [0,1], which is beneficial for convergence at the beginning of the training stage. Since we use periodic boundary conditions, we used random translation and rotations as data-augmentation.

As a result, we obtain 0.008 validation mean squared error, which is slightly higher than the previous result with convolution-based model [3]. Figure 2 provides a qualitative comparison between trained and ground truth spectral responses.

3. Dataset details

In the main dataset we provide hyperspectral images of real and artificial fruits. The miscellaneous class corresponds to fruits and vegetables that do not have a natural counterpart. Approximately 40% of the scenes consist of a single row of objects located at the camera's focal plane. The remaining scenes show two rows of objects, with the focal plane located in between. We keep the position of the white reference panel approximately constant throughout the dataset for easy normalization. The hyperspectral images have a spatial resolution of 512×512 pixels and 204 spectral bands. We also provide an RGB image as seen through the lens of the camera for each scene with the same spatial resolution.

To validate the generalization ability of our framework, we augmented the dataset with 20 additional images in the wild (examples can be seen in Fig. 3). The resulting reconstruction error for these images is 2.54 ± 2.72, a value consistent with the results obtained with the KAUST dataset used to train the encoder.

4. Nanofabrication details

We produce the devices using 15 mm wide and 0.5 mm thick square pieces of fused silica glass as the substrate. Using plasma-enhanced vapor deposition, we deposit a thin layer of amorphous silicon on the glass, the thickness of which is controlled on each sample to match the design requirements. We then spin coat 200 nm of the resist ZEP-520A (ZEON corporation and 40 nm of the resist AR-PC 5090 (ALLRESIST) and pattern the shapes of the nanostructures using an electron beam lithography system with a 100 kV acceleration voltage. Following this, we remove the AR-PC 5090 by submersing each sample for 60 s in deionized water. We develop the samples by submerging them in ZED-50 (ZEON corporation) for 90 s and rinse for 60 s in isopropyl alcohol. We then deposit 22 nm of chromium using electron beam evaporation to create a hard mask and perform liftoff followed by ultrasonic agitation for 1 min. Following this, we remove the unprotected silicon using reactive ion etching, submerge the devices in TechniEtch Cr01 (Microchemicals) for 30 s to remove the metal mask, and rinse with deionized water to obtain the final device.

Figure 3. **Additional samples captured in a real-world setting.**

5. Characterization

We measure the spectral response of our devices in transmission. For accurate characterization, we fabricate each filter separately as a uniform square 500 µm wide. A setup with two 10x microscope objectives allows us to focus broadband (400 nm-1000 nm) linearly polarized light on our samples. A spectrometer then processes the transmitted light. The resulting transmission curves, and their comparison to the theoretical ones, are shown in Fig. 4.

6. Additional results

In this section, we provide additional computational results. Fig. 6 provides additional qualitative comparisons between RGB trained and Hyplex™ model on the segmentation quality on the FVgNET dataset. Fig. 7 illustrates reconstruction efficiency in simulations of Hyplex™ on the samples from KAUST dataset.

7. Real-time processing

The "first layer" of the learning model is purely optical and acquires data at the speed of light. Therefore, the

data acquisition speed of Hyplex™ is limited only by the sensor frame rate (30 FPS in this work). For real-time classification/segmentation tasks, the remaining layers of the network will create delays between the real-time processing of the hyperspectral images and the output for the task. We chose a shallow network implemented in a GPU in this work, which resulted in real-time (> 20 FPS) processing. For better validation purposes, we match the specifications of the dataset we used in training and designed the system to work from $400\,\mathrm{nm}$ to $700\,\mathrm{nm}$ with $10\,\mathrm{nm}$ spectral resolution and 512×512 spatial resolution. In general, our spectral resolution can achieve up to $2\,\mathrm{nm}$, covering the wavelength range from $400\,\mathrm{nm}$ to $700\,\mathrm{nm}$. Using a high-resolution camera sensor currently available in the market (> 12 MP), we could produce a > 2 MP hyperspectral camera with an acquisition speed close to $1\,\mathrm{Tb/s}$. We will provide these additional details in the suppl. material.

.

Validation stats	IoU	F1	Prec	recall	Acc
background	0.9940	0.9970	0.9966	0.9974	0.9948
real potato	0.9626	0.9809	0.9877	0.9743	0.9999
artificial potato	0.9655	0.9824	0.9877	0.9772	0.9999
real apple	0.9522	0.9754	0.9645	0.9867	0.9997
artificial apple	0.9021	0.9485	0.9714	0.9267	0.9988
real orange	0.9786	0.9891	0.9948	0.9836	0.9999
artificial orange	0.9541	0.9764	0.9866	0.9666	0.9996
real grape	0.8294	0.9067	0.8530	0.9677	0.9978
artificial grape	0.8971	0.9457	0.9331	0.9588	0.9990
real lemons	0.8673	0.9289	0.9881	0.8764	0.9992
artificial lemons	0.0009	0.0018	0.7143	0.0009	0.9968
real avocado	0.0365	0.0704	1.0000	0.0365	0.9981
artificial avocado	0.9436	0.9709	0.9764	0.9656	0.9999
real pepper	0.9586	0.9788	0.9808	0.9769	0.9993
artificial pepper	0.9426	0.9704	0.9583	0.9828	0.9996
real unknown	0.9243	0.9606	0.9491	0.9726	0.9980
artificial unknown	0.7016	0.8246	0.7215	0.9621	0.9957
total	0.8124	0.8476	0.9391	0.8537	0.9986
total(-background)	0.8011	0.8382	0.9355	0.8447	0.9988

Table 1. Validation stats on spectral segmentation

Validation stats	IoU	F1	Prec	recall	Acc
background	0.9950	0.9975	0.9968	0.9983	0.9957
real potato	0.9688	0.9841	0.9980	0.9707	0.9999
artificial potato	0.0000	0.0000	0.0000	0.0000	0.9961
real apple	0.9469	0.9727	0.9860	0.9598	0.9993
artificial apple	0.9186	0.9575	0.9301	0.9867	0.9992
real orange	0.9351	0.9664	0.9940	0.9404	0.9994
artificial orange	0.6094	0.7572	0.6239	0.9633	0.9959
real grape	0.0088	0.0174	0.9949	0.0088	0.9917
artificial grape	0.4935	0.6608	0.5015	0.9687	0.9904
real lemons	0.8241	0.9035	0.8387	0.9794	0.9997
artificial lemons	0.0000	0.0000	0.0000	0.0000	0.9973
real avocado	0.5467	0.7069	0.9891	0.5500	0.9992
artificial avocado	0.9708	0.9851	0.9992	0.9716	0.9999
real pepper	0.9068	0.9511	0.9954	0.9107	0.9988
artificial pepper	0.8945	0.9443	0.9081	0.9836	0.9988
real unknown	0.9524	0.9756	0.9762	0.9751	0.9989
artificial unknown	0.7274	0.8421	0.7605	0.9435	0.9967
total	0.6882	0.7425	0.7937	0.7712	0.9975
total(-background)	0.6690	0.7265	0.7810	0.7570	0.9976

Table 2. Validation stats on RGB segmentation

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[2] M. Makarenko, A. Burguete-Lopez, F. Getman, and A. Fratalocchi. Generalized maxwell projections for multimode network photonics. *Scientific Reports*, 10(1):9038, Dec 2020. 1

[3] Maksim Makarenko, Qizhou Wang, Arturo Burguete-Lopez, Fedor Getman, and Andrea Fratalocchi. Robust and scalable flat-optics on flexible substrates via evolutionary neural networks. *Advanced Intelligent Systems*, page 2100105, Aug 2021. 2

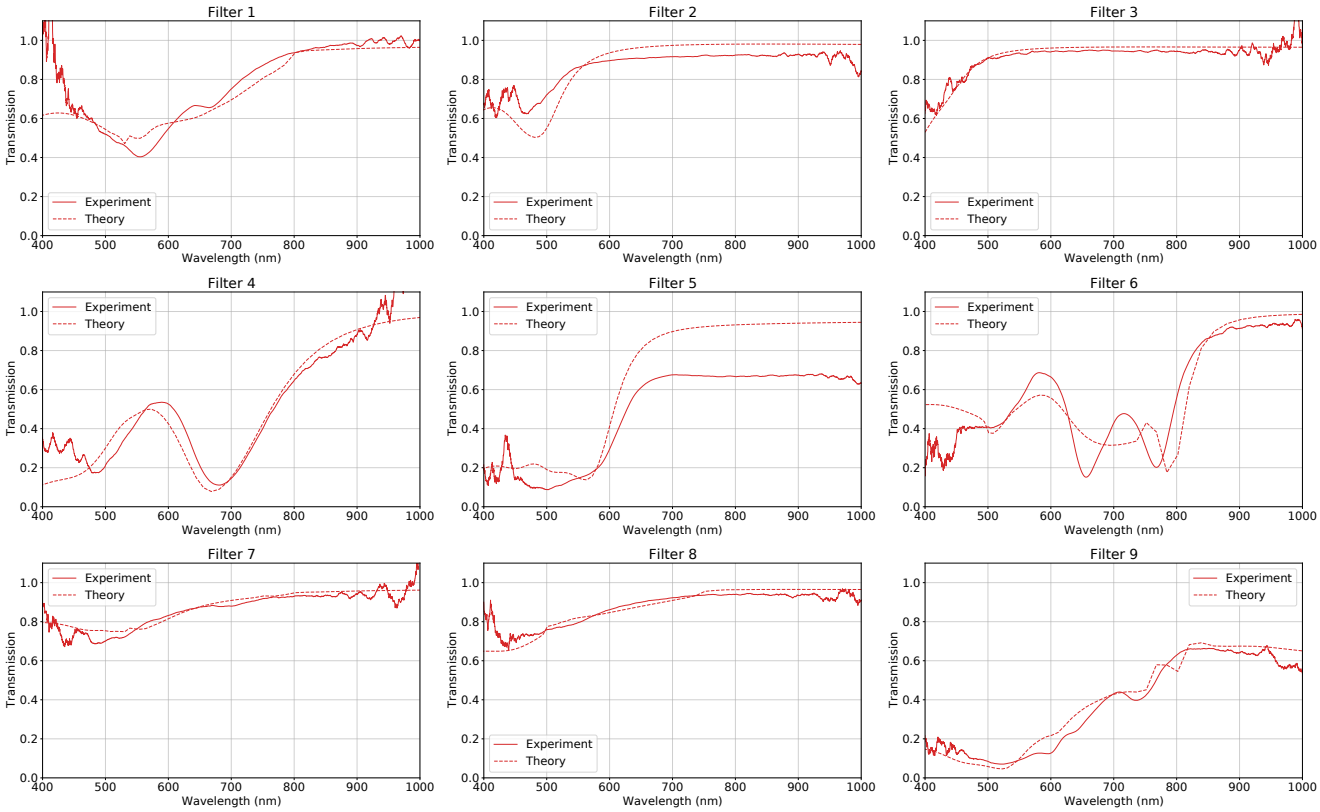

Figure 4. Comparison between theoretical and experimental response of designed filters.

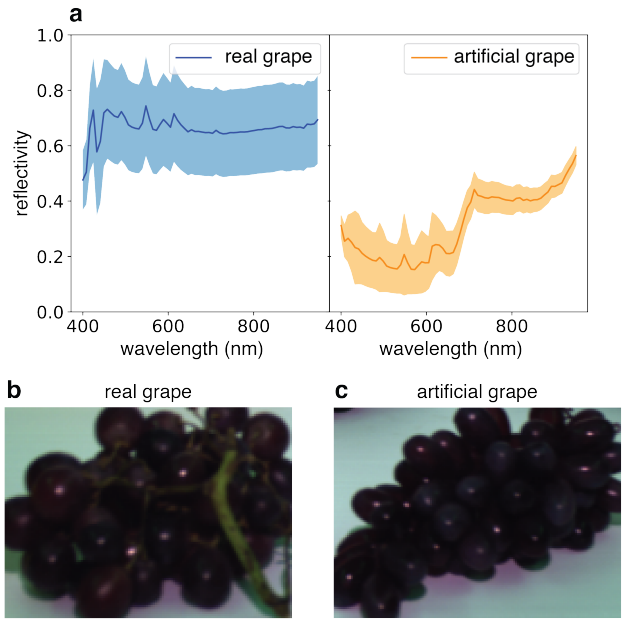

Figure 5. Comparison between spectral responses and RGB images of real and artificial grapes.**a** Reflection spectra of real and artificial grapes. **b** RGB image of real grapes. **c** RGB image of artificial grapes.

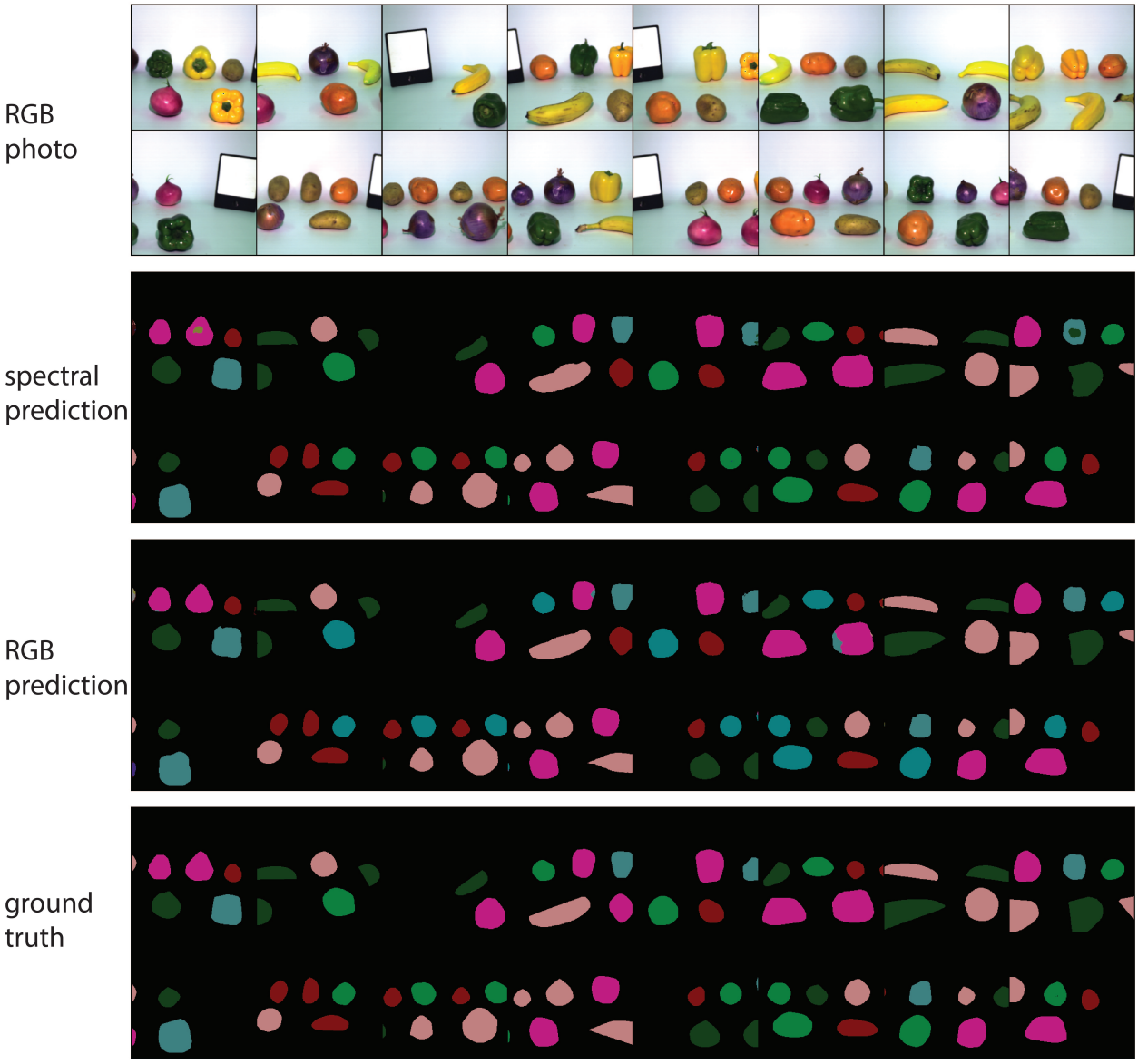

Figure 6. Comparison between RGB and spectral-informed models on semantic fruit segmentation task

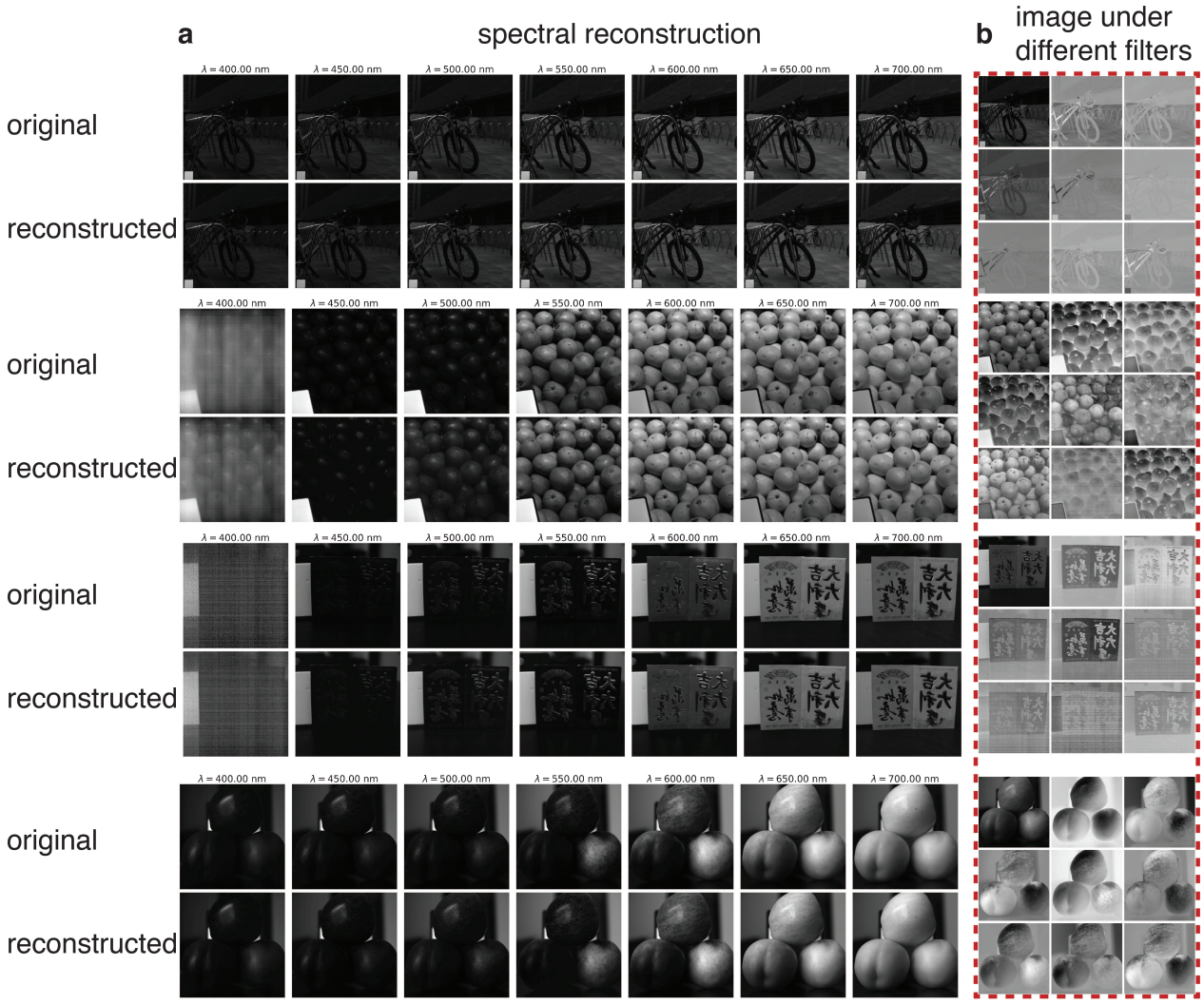

Figure 7. Simulation reconstruction results on KAUST dataset. **a** Image spectral recovery at different wavelengths. **b** Simulated barcode of the scene as it would be perceived by Hyplex™ through each of the nine projectors.