Time-Multiplexed Coded Aperture Imaging: Learned Coded Aperture and Pixel Exposures for Compressive Imaging Systems –Supplemental Material–

Edwin Vargas[1,*], Julien N.P. Martel[2,*], Gordon Wetzstein[2], Henry Arguello[1]
[1]Universidad Industrial de Santander, Colombia [2]Stanford University, USA
edwin.vargas4@correo.uis.edu.co, jnmartel@stanford.edu
gordon.wetzstein@stanford.edu, henarfu@uis.edu.co

1. Derivation of the discrete forward models

In this section, we expose the discrete forward models from the continuous models shown in Equation (8) and (11) of the main paper.

1.1. Compressive Light Field Imaging

We recall the continuous forward model for compressive light field imaging from Equation (8) :

$$e(x) = \int_{\mathcal{V}} \int_{\Delta t} S(x,t')\, l(x,u)\, T(x + s(u-x), t')\, \mathrm{d}u\, \mathrm{d}t'.$$

Considering K discrete time slots, the discretized form of the coded exposure for a given pixel m in the sensor array, can be written as:

$$e_m = \sum_{k=1}^{K} \sum_{\ell=(U-1)/2}^{(U+1)/2} S_m^k\, T_{m+\ell}^k\, l_{m,\ell}, \tag{15}$$

where m, ℓ and K are the indexes for the discretized spatial, angular and time dimensions, and M, U and K are the corresponding number of samples along these dimensions. Note the coded aperture is defined as shifted by a given number of pixels depending on the sub aperture image ℓ.

We define the discrete TMCA as

$$\hat{T}_m = \sum_{k=1}^{K} S_m^k\, T_{m+\ell}^k, \tag{16}$$

and the discrete model of the coded exposure in (15) can be finally expressed as

$$e_m = \sum_{k=1}^{K} \sum_{\ell=(U-1)/2}^{(U+1)/2} \hat{T}_{m,\ell}\, l_{m,\ell}. \tag{17}$$

* denotes equal contributions.

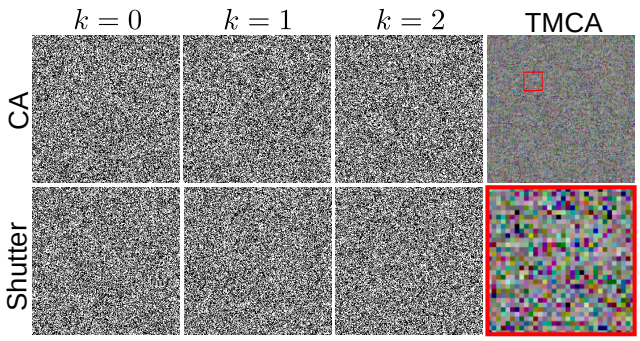

Figure 1. Learned TMCA for compressive spectral imaging. Learned CA and shutter function for the three first time slots $k = 0, 1, 2$. The resultant learned TMCA in CSI is equivalent to a colored coded aperture shown in the last column of the figure.

1.2. Compressive spectral imaging

The continuous forward model for the coded exposure of our compressive spectral imaging system is given in Equation (11) of the main paper:

$$e(x,y) = \int_{\Delta t} S(x,y,t') \iiint T(x',y',\lambda,t')\, I(x',y',\lambda)\, h(x - \mathcal{S}(\lambda) - x', y - y')\, \kappa(\lambda)\, \mathrm{d}x'\, \mathrm{d}y'\, \mathrm{d}\lambda\, \mathrm{d}t'.$$

For a given pixel (m,n) (remember we now index the spatial location in 2D since one of the spatial dimension is "special" and corresponds to the dimension in which the prism disperses light), and considering K discrete time slots, the measurement model can be written as:

$$e_{m,n} = \sum_{k=1}^{K} \sum_{\ell=1}^{L} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} S_{m,n}^k\, f_{i,j,\ell}\, T_{i,j}^k\, h_{m-i,n-j,\ell} \tag{18}$$

where i, j are indices along the discretized spatial dimensions of M, N samples each, ℓ, k denote the indices for the

discretized wavelength and time dimensions, and L, K are the corresponding number of samples along those.

The point spread function (PSF) h corresponds to a propagation model through a unit magnification imaging optics. Further assuming the prism features linear dispersion, the PSF can be expressed as the shifted dirac $\delta_{m-i,n-j,\ell}$. Substituting this expression in Equation (18), and simplifying, the exposure model becomes

$$e_{m,n} = \sum_{k=1}^{K} \sum_{\ell=1}^{L} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} S^k_{i,j+\ell}\, T^k_{i,j}\, f_{i,j,\ell}\, \delta_{m-i,n-j,\ell}, \tag{19}$$

Grouping the time variables in (19), we then define the discrete TMCA:

$$\hat{T}_{i,j,\ell} = \sum_{k=1}^{K} S^k_{i,j+\ell}\, T^k_{i,j}. \tag{20}$$

yielding the coded measurements

$$e_{m,n} = \sum_{\ell=1}^{L} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \hat{T}_{i,j,\ell}\, f_{i,j,\ell}\, \delta_{m-i,n-j,\ell} \tag{21}$$

2. Deriving Equation (12) from Equation (11)

Again, Equation (11) is:

$$e(x,y) = \int_{\Delta t} S(x,y,t') \iiint T(x',y',t')\, I(x',y',\lambda)\, h(x - \mathcal{S}(\lambda) - x', y - y')\, \kappa(\lambda)\, \mathrm{d}x'\, \mathrm{d}y'\, \mathrm{d}\lambda\, \mathrm{d}t'.$$

Since h is the propagation through unit magnification imaging optics and a dispersive element with linear dispersion, the impulse response can be expressed as $h(x - \mathcal{S}(\lambda) - x', y - y') = \delta(x - \lambda - x', y - y')$, and the coded exposure can be simplified as

$$e(x,y) = \int_{\Delta t} \int S(x,y,t')\, T(x - \lambda, y, t')\, I(x - \lambda, y, \lambda)\, \kappa(\lambda)\, \mathrm{d}\lambda\, \mathrm{d}t'. \tag{22}$$

Using the properties of dirac distributions, we express the terms inside the integral as the following convolution

$$S(x,y,t')\, T(x-\lambda, y, t') I(x-\lambda, y, \lambda) = \iint S(x'+\lambda, y', t')\, T(x', y', t')\, I(x', y', \lambda)\, \delta(x - \lambda - x', y - y')\, \mathrm{d}x'\, \mathrm{d}y'. \tag{23}$$

Substituting this expression in Equation (11) , we obtain the following expression (Equation (12) in the main paper) for coded exposure measurements:

$$e(x,y) = \int_{\Delta t} \iiint S(x' + \lambda, y', t')\, T(x', y', t')\, I(x', y', \lambda)\, \delta(x - \lambda - x', y - y')\, \kappa(\lambda)\, \mathrm{d}x'\, \mathrm{d}y'\, \mathrm{d}\lambda\, \mathrm{d}t'. \tag{24}$$

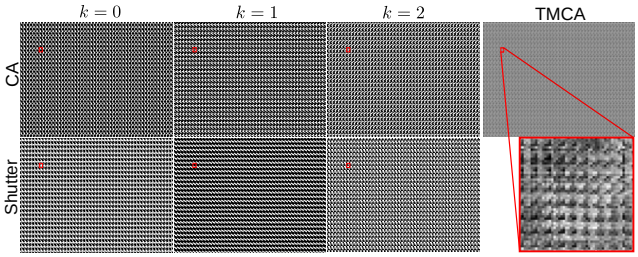

Figure 2. Learned TMCA for compressive light field imaging. Learned CA and shutter function for the first three time slots $k = 0, 1, 2$. We use a 2D array of sub-images for visualization of the TMCA where each sub-image represents the response of the equivalent coded aperture to all rays arriving at one point on the coded aperture from all points on the aperture plane. Thus, the resultant learned TMCA is equivalent to a coded aperture with sensitive angular pixels.

3. Metrics used for spectral imaging

Here we give the formal definitions and some intuition about the metrics we used to evaluate the quality of our compressive spectral imaging system.

- *RMSE:* The root mean square error (RMSE) is a pixel-wise dissimilarity measure between a ground-truth spectral image $\mathbf{x}$ and the reconstructed image $\hat{\mathbf{x}}$ of $N \times M$ pixels and L spectral bands defined as

$$\mathrm{RMSE}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{M \cdot N \cdot L} ||\mathbf{x} - \hat{\mathbf{x}}||_2. \tag{25}$$

- *UIQI:* The universal image quality index (UIQI) was proposed in [5] for evaluating the similarity between two single gray scale images. This metric measure the correlation, contrast and luminance distortion of a reconstructed image with respect to the reference image. The UIQI between two single-band images $\mathbf{a} = [a_1, \cdots, a_{MN}]$ and $\hat{\mathbf{a}} = [\hat{a}_1, \cdots, \hat{a}_{MN}]$ is defined

$$\mathrm{UIQI}(\mathbf{a}, \hat{\mathbf{a}}) = \frac{4\sigma_{a\hat{a}^2}\mu_a\mu_{\hat{a}}}{(\sigma_a^2 + \sigma_{\hat{a}}^2)(\mu_a^2 + \mu_{\hat{a}}^2)} \tag{26}$$

  where $(\mu_a, \mu_{\hat{a}}, \sigma_a^2, \sigma_{\hat{a}}^2)$ are the sample means and variances of $\mathbf{a}$ and $\hat{\mathbf{a}}$, and $\sigma_{a\hat{a}^2}$ is the sample covariance of $(\mathbf{a}, \hat{\mathbf{a}})$. The range of UIQI is $[-1, 1]$ and $\mathrm{UIQI}(\mathbf{a}, \hat{\mathbf{a}}) = 1$ when $\mathbf{a} = \hat{\mathbf{a}}$. Thus, the higher the UIQI, the better spectral reconstruction. Since we work with multi-band images, the overall UIQI metric reported on Table I in the main paper corresponds to the average of the UIQIs over all spectral bands.

- *SAM:* The spectral angle mapper (SAM) was proposed to evaluate the quality of the recovered spectral images by measuring the similarity between reference and estimated spectral signatures [3]. The SAM of two spec-

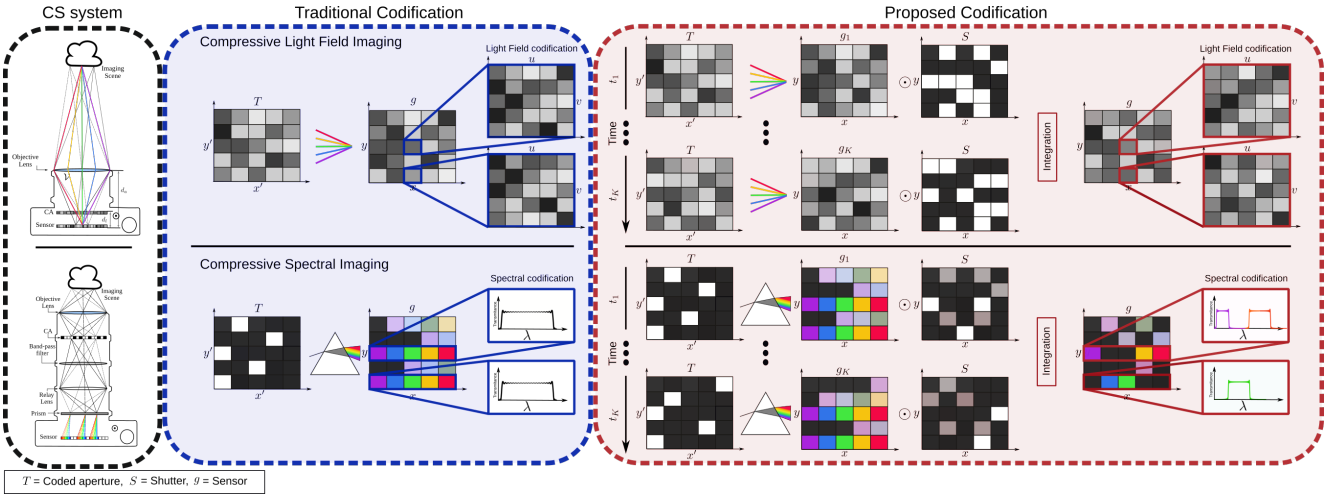

Figure 3. A diagram representing the advantages of the proposed TMCA codification against traditional CA codifications for the two applications we study: compressive light field imaging and compressive hyperspectral imaging.

tral vectors $\mathbf{x}$ and $\hat{\mathbf{x}}$ is defined as

$$\text{SAM}(\mathbf{x}, \hat{\mathbf{x}}) = \arccos\left(\frac{\langle \mathbf{x}, \hat{\mathbf{x}} \rangle}{\|\mathbf{x}\|_2 \|\hat{\mathbf{x}}\|_2}\right). \quad (27)$$

The SAM metric reported in Table I is obtained by averaging the SAMs computed from all $M \cdot N$ image pixels. Since the SAM is an angular quantity, the value of SAM is expressed in degrees and thus belongs to $(-90, 90]$. The smaller the absolute value of SAM, the more higher the spectral similarity is between the recovered image and the ground truth.

- *ERGAS:* The relative dimensionless global error in synthesis (ERGAS) has been proposed to compute the amount of spectral distortion in super resolved spectral images [4]. Here, we employ this quantity to evaluate the recovered spectral images:

$$\text{ERGAS} = 100\sqrt{\frac{1}{L}\sum_{i=0}^{L-1}\left(\frac{\text{RMSE}(\mathbf{x}_i, \hat{\mathbf{x}}_i)}{\mu_i}\right)^2} \quad (28)$$

where μ_i is the mean of the i-th band ($\mathbf{x}_i$) of the spectral image, and L is the number of spectral bands. The smaller ERGAS, the smaller the spectral distortion.

- *DD:* The final metric reported in Table I is the degree of distortion (DD) between two spectral images which is defined as

$$\text{DD}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{N \cdot M \cdot L}\|\mathbf{x} - \hat{\mathbf{x}}\|_1. \quad (29)$$

The smaller DD, the better the recovered spectral image.

4. Simulation details

In this section we present additional qualitative results of the simulations in compressive hyperspectral imaging and compressive light field imaging.

4.1. Coded Hyperspectral Imaging

Training details: We learn the end-to-end model for hyperspectral imaging using 160 spectral images from the ICVL dataset [1]. We used cropped images at a size of 256×256 with $L = 12$ spectral bands. The L spectral bands corresponds to the following wavelengths in nm: $[480, 500, 510, 530, 550, 560, 570, 590, 600, 620, 640, 650]$. We set the number of time slots in the TMCA encoder to $K = 8$. The U-Net is trained for 500 epochs using ADAM optimizer. We applied a learning rate decay of factor 0.5 every 150 epoch with an initial rate of 0.0001. We display the coded apertures and shutter function learned in our pipeline in Figure 1.

Rational behind our baselines, we compare the proposed TMCA codification against four different baselines: a) the traditional CASSI codification using random binary patterns and reconstructed using the alternative direction method of multipliers (ADMM) b) the CASSI codification using a trained U-Net as a decoder c) the proposed TMCA codification and reconstruction pipeline using random (non-optimized) codes d) the CASSI system jointly learning the codification and the U-Net as a decoder and e) our full TMCA codification with learned codes.

Baseline a) shows the performance of a system that does not use our codification, and uses a conventional optimization technique as a decoder. The baseline b) still uses the traditional codification but now uses a modern NN decoder,

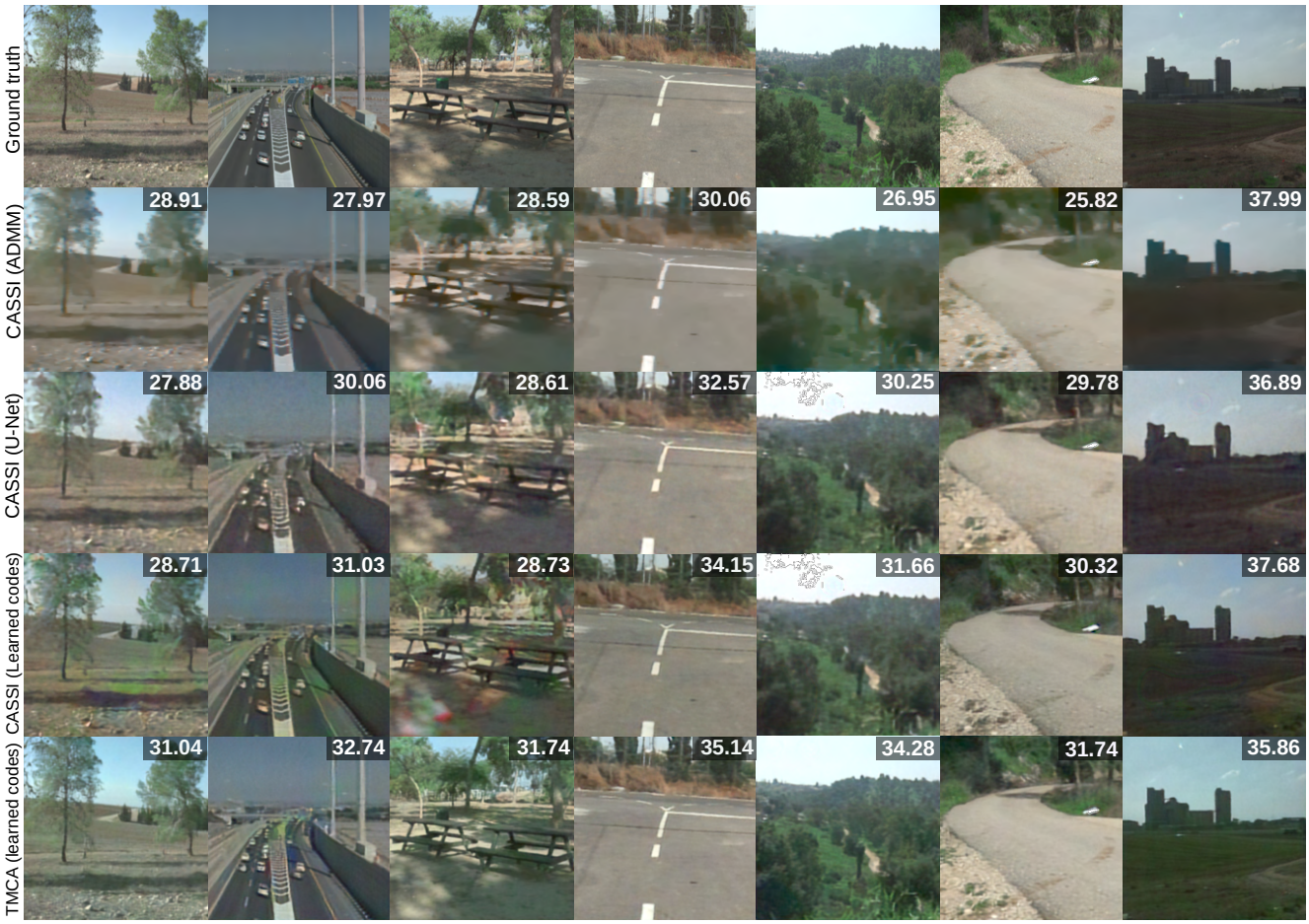

Figure 4. Additional results for compressive spectral imaging on the ICVL 1 test fold comparing our TMCA (fifth row) against the CASSI using a conventional optimization technique (ADMM) as a decoder (second row), a U-Net decoder (third row), and jointly learned codes and U-Net decoder (fourth row).

thus allowing us to state that the results we observe with baseline c) that uses our TMCA codification cannot only be attributed to the fact we use a NN decoder, but indeed that the TMCA codification is better. The baseline (d) employs joint optimized codes from the CASSI system and trained NN confirming that the gain of the proposed TMCA codification does not solely depend on the end-to-end optimization of the optics and reconstruction algorithm. Finally, d) shows that optimizing the TMCA codification itself is also beneficial.

Additional qualitative results We show qualitative results for a few more reconstructed hyperspectral images in Figure 4.

Mapping spectral bands to RGB Our hyperspectral images are mapped to an RGB composite image by selecting three spectral channels corresponding to wavelengths of 650,550 and 480 nm.

Spectral signatures We show qualitative and quantitative comparisons of the full spectral signatures ($L = 12$) for two different points taken in a randomly sampled image of the ICVL 1 dataset in Figure 5 and show qualitatively $L = 6$ bands for two other randomly sampled images in Figure 8.

Sensitivity to noise and shutter length We performed additional experiments varying the Gaussian readout noise for a constant shutter length ($K = 8$) and varying the shutter length for a fixed level of noise. The results can be found in Table 1 and 2 showing the less noise the better. We find an optimal shutter length value to be $K = 16$.

$\sigma^2 =$	10^{-4}	10^{-3}	10^{-2}	10^{-1}
PSNR	32.051	31.723	31.311	30.207
UIQI	0.979	0.976	0.977	0.965
SAM	5.43	5.73	5.98	6.86
ERGAS	12.57	12.92	13.49	15.41
DD	0.017	0.018	0.019	0.022

Table 1. Adding additive gaussian noise (ICVL 1 dataset)

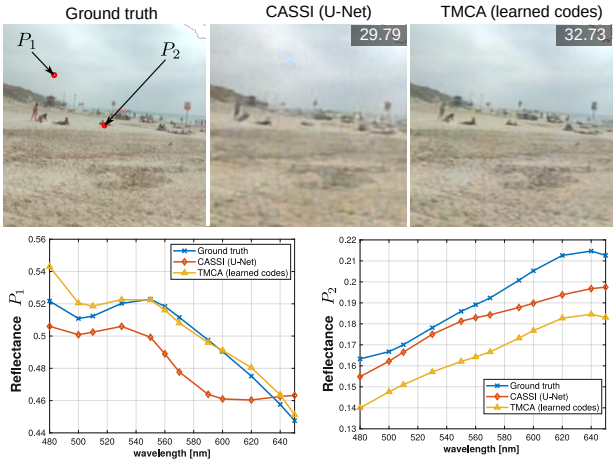

Figure 5. Spectral signatures for two points in an image randomly sampled from the ICVL 1 dataset. The $L = 12$ spectral bands are shown for a point in the sky and a point in the sand on the reconstructions performed by our method using TMCA and learned codes and the conventional CASSI with a U-Net.

$K =$	2	4	8	16
PSNR	31.032	30.937	31.723	31.858
UIQI	0.971	0.972	0.976	0.978
SAM	6.29	6.21	5.73	5.54
ERGAS	14.17	14.04	12.92	12.63
DD	0.021	0.020	0.018	0.018

Table 2. Varying the shutter length (ICVL 1 dataset)

4.2. Compressive Light Field Imaging

Training details: We learn our end-to-end model for compressive light field imaging by using the aggregate dataset described in the main paper. For, this experiment we aim to recover light fields with 5×5 angular views and resolution of 480×270 to match with the spatial resolution of the experimental setup. We also set the number of time slots in the TMCA encoder to $K = 8$. We use randomly cropped patches of those images of spatial size 11×11 for training, the decoder is the deep spatial-angular convolutional sub-network proposed in [2] which is trained for 500 epochs using ADAM optimizer. Similar to the spectral application, we applied a learning rate decay of factor 0.5 every 150 epochs with an initial rate of 0.0001. We display the learned coded apertures and shutter function in our pipeline in Figure 2. We use a 2D array of sub-images for visualization of the TMCA where each sub-image represents the response of the equivalent coded aperture to all rays arriving at one point on the coded aperture from all points on the aperture plane.

Additional qualitative results We show qualitative results for the central view of four additional reconstructed light fields in Figure 9 and show the 5×5 reconstructed angular views of yet another light field from the Lytro dataset in Figure 10.

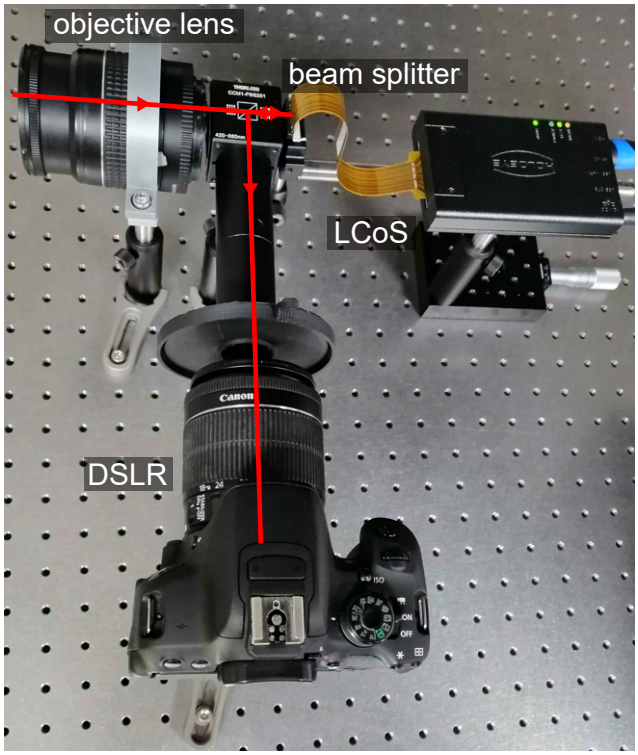

Figure 6. A photograph of the optical setup used in our compressive light field imaging experiments. The light path is shown in red and the various components are labelled (detailed components' description is found in the main paper).

Rational behind our baselines The rational behind our baselines for compressive light field imaging is the same as for hyperspectral imaging. Baseline a) uses the traditional codification and an ADMM decoder while b) swaps the ADMM for a deep neural network [2] and shows the NN alone does not explain the better results obtained in baseline c) with our TMCA codification with random codes. Baseline d) jointly optimizes traditional codification with a deep decoder shows that the improvement of e) our TMCA codification with learned codification cannot be only attributed to the end-to-end optimization.

Sensitivity to noise and shutter length We show quantitative results for an additional experiments varying the Gaussian readout noise for a constant shutter length ($K = 8$) and varying the shutter length for a fixed level of noise. The results can be find in Table 3 and 4 showing the less noise the better as expected and an optimal shutter length value for $K = 8$.

$\sigma^2 =$	10^{-4}	10^{-3}	10^{-2}	10^{-1}
PSNR	34.67	33.57	31.65	28.98
SSIM	0.923	0.911	0.851	0.772

Table 3. Adding additive Gaussian noise (Light-field Lytro dataset)

$k =$	2	4	8	16
PSNR	34.11	34.66	34.67	34.47
SSIM	0.894	0.908	0.911	0.916

Table 4. Varying the shutter length (Light-field Lytro dataset)

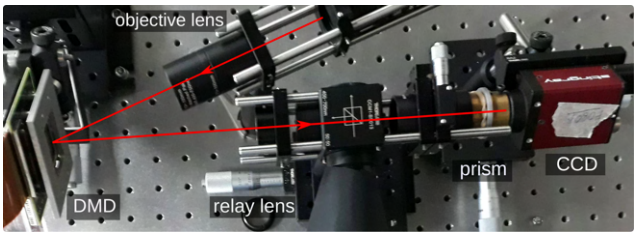

Figure 7. A photograph of the optical setup implementing our hyperspectral compressive imaging system. The light path is shown in red and the various components are labelled (a detailed components' description is found in the main paper.)

5. Optical setups

Compressive light field imaging The setup consists of an objective lens projecting the image on a LCoS imaged with a DSLR (equipped with its objective lens) through a beamsplitter. The component details are given in the main paper. A photography of the setup used in our experiments in shown in Figure 6.

Compressive hyperspectral imaging The setup consists of an objective lens projecting the image on a DMD imaged with a monochromatic CCD through a relay lens and prism dispersing light. The component details are given in the main paper. A photography of the setup used in our experiments in shown in Figure 7.

Acknowledgments

J.N.P.M was supported by a Swiss National Foundation (SNF) Fellowship (P2EZP2_181817), G.W. was supported by an NSF Award (1839974), and a PECASE by the ARL. H.A. was supported by the Fullbright 2019 Visiting Scholar Program.

References

[1] Boaz Arad and Ohad Ben-Shahar. Sparse recovery of hyperspectral signal from natural rgb images. In *European Conference on Computer Vision*, pages 19–34. Springer, 2016. 3

[2] Mantang Guo, Junhui Hou, Jing Jin, Jie Chen, and Lap-Pui Chau. Deep spatial-angular regularization for compressive light field reconstruction over coded apertures. In *European Conference on Computer Vision*, pages 278–294. Springer, 2020. 5, 8

[3] Fred A Kruse, AB Lefkoff, JW Boardman, KB Heidebrecht, AT Shapiro, PJ Barloon, and AFH Goetz. The spectral image processing system (sips)—interactive visualization and analysis of imaging spectrometer data. *Remote sensing of environment*, 44(2-3):145–163, 1993. 2

[4] Lucien Wald. Quality of high resolution synthesised images: Is there a simple criterion? In *Third conference" Fusion of Earth data: merging point measurements, raster maps and remotely sensed images"*, pages 99–103. SEE/URISCA, 2000. 3

[5] Zhou Wang and Alan C Bovik. A universal image quality index. *IEEE signal processing letters*, 9(3):81–84, 2002. 2

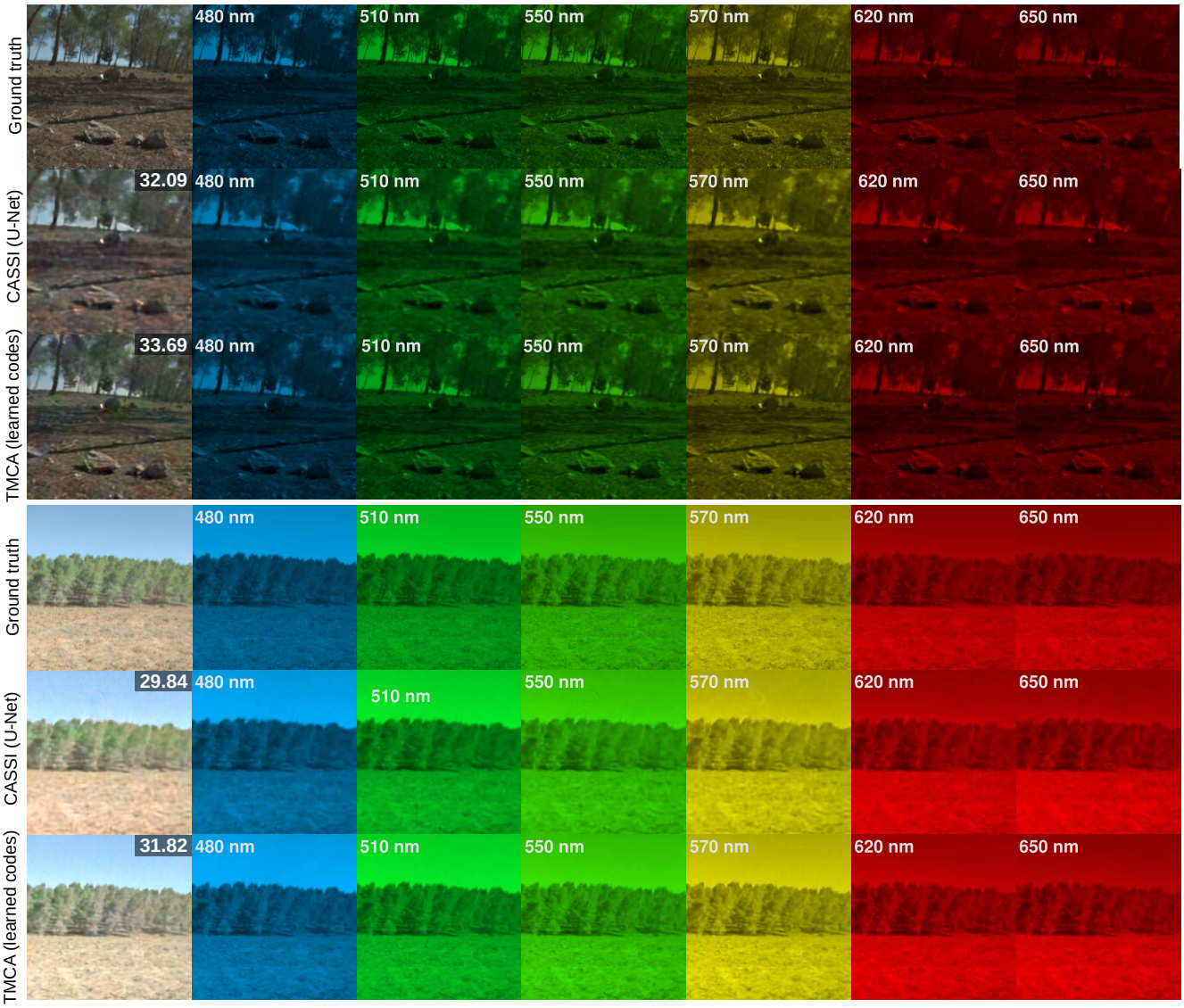

Figure 8. Results of our compressive spectral imaging reconstruction on two synthetic images of the ICVL 1 Dataset showing 6 different spectral bands (out of $L = 12$).

Figure 9. Additional light field results comparing our codification with TMCA using the deep network [2] as a decoder against (fourth row) the CLFP baselines with a sparse dictionnary coding method as a decoder (second row) and the same codification also using the deep network architecture from [2] (third row). Numbers in the top right corner indicate PSNR compared to the ground truth (first row).

Figure 10. The 5×5 angular views reconstructed from a randomly sampled light field of the Lytro dataset comparing our method with the CLFP baseline.