Deep Material-aware Cross-spectral Stereo Matching Supplementary Material

Tiancheng Zhi, Bernardo R. Pires, Martial Hebert, and Srinivasa G. Narasimhan
Carnegie Mellon University
{tzhi,bpires,hebert,srinivas}@cs.cmu.edu

1. Discussion about the Symmetric CNN

This section discusses how the use of a symmetric convolutional neural network (CNN) prevents the spectral translation network (STN) from learning disparity.

1.1. Problem Simplification and Basic Property

Because white balance and exposure correction are global operations, they can be ignored when considering the displacement. Let I be the input image, $\mathcal{F}$ be the filter predicted by the symmetric CNN and O be the output image. Without loss of generality assume I, $\mathcal{F}$ and O to be single channel. Formally,

$$O(x, y) = I(x, y)\mathcal{F}(x, y) \tag{1}$$

Now consider the left-right flipped image (for simplicity assume image coordinate system is centered on the image):

$$I'(x, y) = I(-x, y) \tag{2}$$

Because the filtering kernels in the symmetric CNN are left-right symmetric, a flipped input image $I'(x, y) = I(-x, y)$ leads to a flipped output filter $\mathcal{F}'(x, y) = \mathcal{F}(-x, y)$. Thus the output O' of the flipped image I' is:

$$O'(x, y) = I(-x, y)\mathcal{F}(-x, y) = O(-x, y) \tag{3}$$

1.2. Ignoring Spectral Difference

Spectral difference is ignored here and we focus on geometric difference (disparity) first.

We assume that if STN learns disparity, it only learns non-negative disparity. This is a reasonable assumption because the training data obey the epipolar constraint, which allows only non-negative disparity.

The STN does not learn disparity because it cannot shift the image I to get the image O. We explain why the STN does not shift I to get O by contradiction.

Assume that the STN shifts I to get O according to disparity map $\Delta(x, y) \geq 0$, *i.e.*,

$$O(x, y) = I(x + \Delta(x, y), y) \tag{4}$$

Similarly, assume that the STN shifts I' to get O' according to disparity map $\Delta'(x, y) \geq 0$, *i.e.*,

$$O'(x, y) = I'(x + \Delta'(x, y), y) \tag{5}$$

$\Delta(x, y)$ and $\Delta'(x, y)$ are assumed to be smooth.

Start with Equation 3:

$$O'(x, y) = O(-x, y) \tag{6}$$

Now apply Equation 5 to left hand side and Equation 4 to right hand side to get:

$$I'(x + \Delta'(x, y), y) = I(-x + \Delta(-x, y), y) \tag{7}$$

Now flipping the right hand side gives:

$$I'(x + \Delta'(x, y), y) = I'(x - \Delta(-x, y), y) \tag{8}$$

Since $\Delta'(x, y) \geq 0$ and $\Delta(-x, y) \geq 0$ and they are smooth as assumed, $\Delta'(x, y)$ and $\Delta(-x, y)$ must be 0 assuming there are no duplicated patches in the image, and therefore the STN cannot shift the image.

1.3. Considering Spectral Difference

We further consider spectral difference by introducing $\Lambda(x, y)$ and $\Lambda'(x, y)$ as spectral translation factors and extend Equation 4 and 5 to:

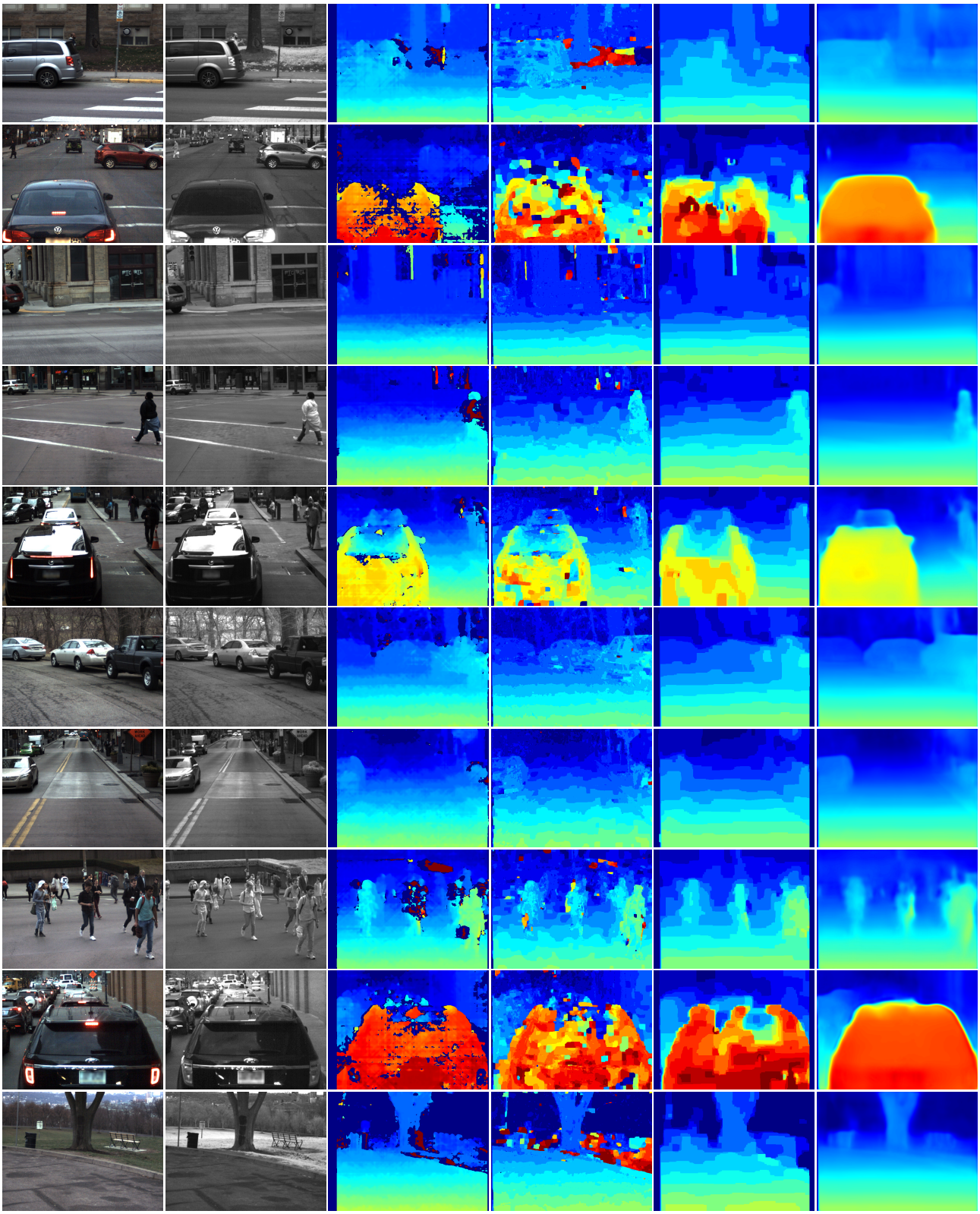
$$O(x, y) = \Lambda(x, y)I(x + \Delta(x, y), y) \tag{9}$$

$$O'(x, y) = \Lambda'(x, y)I(x + \Delta'(x, y), y) \tag{10}$$

We assume that flipping I leads to the flipped $\Lambda(x, y)$, because flipping does not change spectral property. Formally,

$$\Lambda'(x, y) = \Lambda(-x, y) \tag{11}$$

Then we get the same Equation 7 because $\Lambda'(x, y)$ and $\Lambda(-x, y)$ are canceled out. Finally we can make the same conclusion that STN does not shift the input image and thus does not learn disparity.



(a) Left RGB (b) Right NIR (c) CMA [1] (d) ANCC [2] (e) DASC [3] (f) Proposed

Figure 1. More qualitative comparisons. The proposed method provides less noisy disparity maps and performs better on lights, glass and glossy surfaces.

Method	Common		Light		Glass		Glossy		Vegetation		Skin		Clothing		Bag		Mean	
	>3	>5	>3	>5	>3	>5	>3	>5	>3	>5	>3	>5	>3	>5	>3	>5	>3	>5
CMA [1]	2.00	1.25	21.38	12.47	7.22	4.05	14.44	11.70	7.09	4.39	7.00	3.70	19.30	10.53	18.04	8.25	12.06	7.04
ANCC [2]	1.30	1.02	5.79	3.56	8.80	3.52	8.59	6.40	21.62	17.91	4.53	3.70	5.96	4.91	6.19	3.61	7.85	5.58
DASC [3]	0.98	0.46	2.90	0.89	5.81	1.94	8.59	4.57	2.53	0.68	4.12	2.06	**0.70**	**0.00**	5.15	1.03	3.85	1.45
Proposed	**0.00**	**0.00**	**0.45**	**0.00**	**0.35**	**0.00**	**0.18**	**0.00**	**0.00**	**0.00**	**1.65**	**0.00**	**0.70**	0.35	**0.00**	**0.00**	**0.42**	**0.04**

Table 1. Another evaluation metric. Disparity BPR with threshold 3 and 5 pixels is reported in percentage for each material. Our method outperforms other methods generally. DASC [3] performs better than our method on clothing, possibly due to the weak relationship between RGB and NIR appearances of clothing.

2. Preprocessing in Experiments

Preprocessing is done before sending images into the networks. Black level correction and normalization are applied to make images linear and have mean pixel value of 0.5. Exposure times used in the STN are adjusted according to the normalization scale. Pixels are clamped within $[0, 5]$.

Let I be the original image. The image after black level correction is:

$$I_b = \frac{\max\{I - b, 0\}}{255 - b} \tag{12}$$

where $b = 2$ is the maximum pixel value when there is no light.

The image after normalization is

$$I_n = \min\{\frac{0.5 I_b}{mean(I_b) + \epsilon}, 5\} \tag{13}$$

where $\epsilon = 0.001$ to avoid the division by zero.

The exposure time Δt is corrected to be:

$$\Delta t_n = \frac{0.5 \Delta t}{mean(I_b) + \epsilon} \tag{14}$$

where $\epsilon = 0.001$ to avoid the division by zero.

3. More Qualitative Results

See Figure 1 for more qualitative comparisons.

4. Another Evaluation Metric

We report the bad pixel rate (BPR) as another evaluation metric in Table 1. Our method outperforms other methods generally. DASC [3] performs better on clothing, possibly because of the weak relationship between clothing's RGB and NIR appearances.

5. Discussion about the Intermediate Results

As shown in Figure 3 (d), (e) and (f) and pointed out by the caption, some clothing and bags fail in spectral translation. It is probably because of the weak relationship between its RGB and NIR intensities. A typical example is that a dark clothing in RGB may look dark or bright in NIR. Thus it is hard for STN to predict the correct intensity translation in this case.

However, the structural dissimilarity term in the alignment loss can partially solve this problem. Spectral translation usually preserves the structure. In this case, the corresponding regions could be matched because of high structural similarity. For example, the edge of clothing may be matched in disparity prediction.

References

[1] W. W.-C. Chiu, U. Blanke, and M. Fritz. Improving the kinect by cross-modal stereo. In *BMVC*, 2011. 2, 3

[2] Y. S. Heo, K. M. Lee, and S. U. Lee. Robust stereo matching using adaptive normalized cross-correlation. *TPAMI*, 2011. 2, 3

[3] S. Kim, D. Min, B. Ham, S. Ryu, M. N. Do, and K. Sohn. Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *CVPR*, 2015. 2, 3