Deep Material-aware Cross-spectral Stereo Matching

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

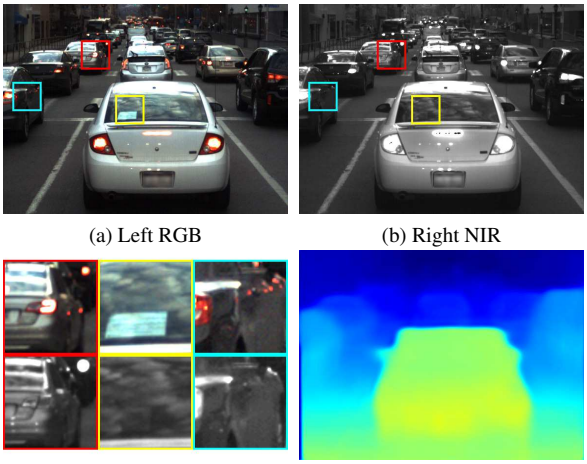
Abstract

*Cross-spectral imaging provides strong benefits for recognition and detection tasks. Often, multiple cameras are used for cross-spectral imaging, thus requiring image alignment, or disparity estimation in a stereo setting. Increasingly, multi-camera cross-spectral systems are embedded in active RGBD devices (*e.g. *RGB-NIR cameras in Kinect and iPhone X). Hence, stereo matching also provides an opportunity to obtain depth without an active projector source. However, matching images from different spectral bands is challenging because of large appearance variations. We develop a novel deep learning framework to simultaneously transform images across spectral bands and estimate disparity. A material-aware loss function is incorporated within the disparity prediction network to handle regions with unreliable matching such as light sources, glass windshields and glossy surfaces. No depth supervision is required by our method. To evaluate our method, we used a vehicle-mounted RGB-NIR stereo system to collect 13.7 hours of video data across a range of areas in and around a city. Experiments show that our method achieves strong performance and reaches real-time speed.*

1. Introduction

Cross-spectral imaging is broadly used in computer vision and image processing. Near infrared (NIR), short-wave infrared (SWIR) and mid-wave infrared (MWIR) images assist RGB images in face recognition [1, 16, 23, 29]. RGB-NIR pairs are utilized for shadow detection [35], scene recognition [2] and scene parsing [5]. NIR images also help color image enhancement [42] and dehazing [11]. Blue fluorescence and ultraviolet images assist skin appearance modeling [24]. Color-thermal images help pedestrian detection [19, 40].

As multi-camera multi-spectral systems become more common in modern devices (*e.g.* RGB-NIR cameras in iPhone X and Kinect), the cross-spectral alignment problem is becoming critical since most cross-spectral algorithms require aligned images as input. Aligning images in hardware with a beam splitter is often impractical as it leads to significant light loss and thus needs longer exposure, resulting in motion blur. Stereo matching handles this problem by estimating disparity from a rectified image pair. Aligned images are obtained by image warping according to disparity. Stereo matching also provides an opportunity to obtain depth without an active projector source (as is done in the Kinect), helping tasks like detection [14] and tracking [37].



(a) Left RGB (b) Right NIR (c) Difficult regions for matching (d) Predicted disparity

Figure 1. A challenging case for RGB-NIR stereo matching and our result. Red box: The light source is visible in RGB but not in NIR. Yellow box: The transmittance and reflectance of the windshield are different in RGB and NIR. Cyan box (brightened): Some light sources reflected by the specular car surface are only visible in RGB. Our approach uses a deep learning based simultaneous disparity prediction and spectral translation technique with material-aware confidence assessment to perform this challenging task.

Cross-spectral stereo matching is challenging because of large appearance changes in different spectra. Figure 1 is an example of RGB-NIR stereo. Headlights have different apparent sizes or intensities in RGB and NIR. LED tail-lights are not visible in NIR. Glass often shows different light transmittance and reflectance in RGB and NIR. Glossy surfaces have different specular reflectance. Additionally, veg-

etation and clothing often show large spectral difference.

In this paper, we propose a deep learning based RGB-NIR stereo matching method without depth supervision. We use two networks to simultaneously predict disparity and remove the spectral difference. A disparity prediction network (DPN) estimates disparity maps based on a RGB-NIR stereo pair, and a spectral translation network (STN) converts a RGB image into a pseudo-NIR image. The losses are constructed by reprojecting and matching the NIR and the pseudo-NIR images, thus both the geometric and spectral differences are encoded. To make sure the disparity is only learned by the DPN, we use a symmetric network design to prevent the STN from learning geometric differences.

Though the DPN and STN work well in many cases, certain materials cannot be handled correctly due to unreliable matching. 'Unreliable' means it is hard to find good matches due to large spectral difference, or the matches found correspond to incorrect disparities (*e.g*. matches on reflections). As shown in Figure 1 and 4, light sources in RGB may be absent in NIR, or show different apparent sizes resulting in incorrect matches. The transmitted and reflected scenes on glass and specular reflection on glossy surfaces may be matched but do not represent the real disparity. These are fundamental problems occurring often and cannot be ignored. We address the problems by using a material recognition network to identify unreliable regions and inferring their disparities from the context. The DPN loss assesses pixel confidence according to the material probability and the predicted disparity, and utilizes a confidence-weighted smoothing technique to backpropagate more gradients to lower confidence pixels. This method significantly improves results in unreliable regions.

We have collected 13.7 hours of RGB-NIR stereo frames covering different scenes, lighting conditions and materials. The images were captured from a vehicle driven in and around a city. Challenging cases for matching appear very frequently, including lights, windshields, glossy surfaces, clothing and vegetation. We labeled material segments on a subset of the images to train the aforementioned material recognition network. Additionally, we labeled sparse disparities on a test subset for evaluation. To our knowledge, this is the first outdoor RGB-NIR stereo dataset with a large range of challenging materials at this scale. We experimented on this specific but important domain of driving in an urban environment and will extend it to indoor or other outdoor domains in the future. Experimental results show that the proposed method outperforms other comparable methods and reaches real-time speed. This method could be extended to other spectra like SWIR or thermal.

2. Related Work

Cross-modal Stereo Matching: The key to cross-modal stereo matching is to compute an invariant between different imaging modalities. Chiu *et al*. [4] proposed a cross-modal adaptation method via linear channel combination. Heo *et al*. [17] presented a similarity measure robust to varying illumination and color. Heo *et al*. [18] also proposed a method to jointly produce color consistent stereo images and disparity under radiometric variation. Pinggera *et al*. [34] showed that the HOG [7] feature helps visible-thermal matching. Shen *et al*. [36] proposed a two-phase scheme with robust selective normalized cross correlation. Kim *et al*. [25] designed a descriptor based on self-similarity and extended it into a deep learning version [26]. Jeon *et al*. [22] presented a color-monochrome matching method in low-light conditions by compensating for the radiometric differences. These methods are based on feature or region matching without material awareness and are unreliable for materials such as lights, glass or glossy surfaces.

Unsupervised Deep Depth Estimation: Unsupervised depth estimation CNNs are usually trained with a smoothness prior and reprojection error. Garg *et al*. [12] proposed a monocular method with Taylor expansion and coarse-to-fine training. Godard *et al*. [13] presented a monocular depth network with left-right consistency. Zhou *et al*. [44] proposed a structure from motion network to predict depth and camera pose. Zhou *et al*. [43] presented a stereo matching method by selecting confident matches and training data. Tonioni *et al*. [38] showed that deep stereo matching models can be fine-tuned with the output of conventional stereo algorithms. All these methods deal with only RGB images rather than cross-spectral images, with no consideration for difficult non-Lambertian materials.

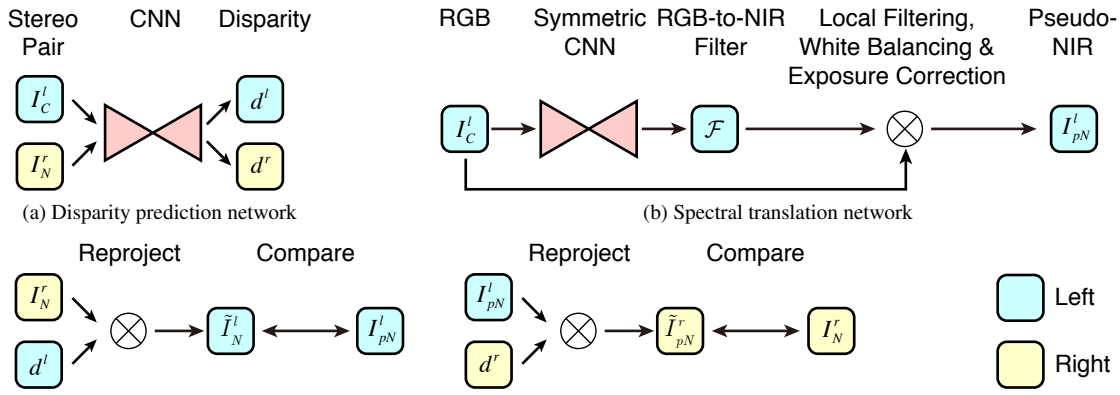
3. Simultaneous Disparity Prediction and Spectral Translation

To compensate for appearance differences between RGB and NIR and extract disparity, we present an unsupervised scheme that trains two networks simultaneously to respectively learn disparity and spectral translation with reprojection error (Figure 2).

3.1. Model Overview

Our approach consists of a disparity prediction network (DPN) and a spectral translation network (STN). The DPN design follows Godard *et al*. [13] but the input is replaced with a RGB-NIR stereo pair $\{I_C^l, I_N^r\}$, where superscripts l and r refer to left and right images. Left-right disparity maps $\{d^l, d^r\}$ are predicted by DPN. STN translates a RGB image I_C^l into a pseudo-NIR image I_{pN}^l. Translation from NIR to RGB is not used because it is hard to add information to a 1-channel image to create a 3-channel image.

Both networks use reprojection error as main loss. Given the right NIR image I_N^r and the left disparity d^l, we reproject the left NIR image $\tilde{I}_N^l$ via differentiable warping [21], similar to previous works [13, 28, 44]. Let



(a) Disparity prediction network (b) Spectral translation network

(c) Loss construction: comparing aligned NIR and pseudo-NIR images (d) Legend

Figure 2. Model overview. The disparity prediction network (DPN) predicts left-right disparity for a RGB-NIR stereo input. The spectral translation network (STN) converts the left RGB image into a pseudo-NIR image. The two networks are trained simultaneously with reprojection error. The symmetric CNN in (b) prevents the STN learning disparity.

$\omega(I, d)$ be the operator warping I according to disparity d, *i.e.*, $\omega(I,d)(x,y) = I(x + d(x,y), y)$. Then $\tilde{I}^l_N = \omega(I^r_N, -d^l)$. Symmetrically, the warped pseudo-NIR image $\tilde{I}^r_{pN} = \omega(I^l_{pN}, d^r)$. Error is calculated between the warped NIR image $\tilde{I}^l_N$ and the pseudo-NIR image I^l_{pN}, and the warped pseudo-NIR image $\tilde{I}^r_{pN}$ and the NIR image I^r_N.

3.2. Disparity Prediction Network

The DPN predicts left-right disparities $\{d^l, d^r\}$ based on a RGB-NIR stereo pair $\{I^l_C, I^r_N\}$. The network structure proposed by Godard *et al*. [13] is adopted. Convolutional layers are followed by batch normalization [20] (except for output layers) and ELU [6] activation. The output disparity is scaled by factor η for a good initialization. The loss has a view consistency term L_v, an alignment term L_a and a smoothness term L_s following Godard *et al*. [13].

$$L_{DPN} = \lambda_v(L^l_v + L^r_v) + \lambda_a(L^l_a + L^r_a) + \lambda_s(L^l_s + L^r_s) \quad (1)$$

For simplicity, only the left terms are described below and the right ones can be derived similarly. Multi-scale prediction is done by adding similar loss functions at four scales.

The view consistency term L^l_v describes the consistency of left-right disparity maps. Let N be the number of the pixels in one image, and Ω be the pixel coordinate space.

$$L^l_v = \frac{1}{N}\sum_{\mathbf{p}\in\Omega} |d^l(\mathbf{p}) - \omega(d^r, -d^l)(\mathbf{p})| \quad (2)$$

The alignment term L^l_a compares intensity and structure between aligned NIR and pseudo-NIR images. Let $\delta(I_1, I_2)$ be the structural dissimilarity function [39]. Then,

$$L^l_a = \frac{1}{N}\sum_{\mathbf{p}\in\Omega} (\alpha\delta(I^l_{pN}, \tilde{I}^l_N)(\mathbf{p}) + (1-\alpha)|I^l_{pN}(\mathbf{p}) - \tilde{I}^l_N(\mathbf{p})|) \quad (3)$$

where α is set to be 0.85 as suggested by Godard *et al*. [13].

The smoothness term L^l_s is edge-aware to allow noncontinuous disparity at image edges:

$$L^l_s = \frac{1}{N}\sum_{\mathbf{p}\in\Omega} ((\left|\frac{\partial d^l}{\partial x}\right| e^{-|S_x * I^l_C|} + \left|\frac{\partial d^l}{\partial y}\right| e^{-|S_y * I^l_C|})(\mathbf{p})) \quad (4)$$

where S_x and S_y are Sobel operators and the filtered RGB channels are averaged into one channel.

3.3. Spectral Translation Network

The RGB-NIR cameras are radiometrically calibrated and their varying white balancing gains (g_R for red and g_B for blue) and exposure times Δt_C and Δt_N are known. The spectral translation network (STN) converts a RGB image I^l_C into a pseudo-NIR image I^l_{pN} via local filtering, white balancing, and exposure correction (Figure 2). Let $\mathcal{G}_{\boldsymbol{\theta}_1}$ be the white balancing operator with learnable parameter $\boldsymbol{\theta}_1$, and $\mathcal{F}^{(\mathbf{p})}_{\boldsymbol{\theta}_2}$ be the filtering operation with predicted parameter $\boldsymbol{\theta}_2$ for each position $\mathbf{p}$. The pseudo-NIR image is:

$$I^l_{pN}(\mathbf{p}) = \frac{\Delta t_N}{\Delta t_C}\mathcal{G}_{\boldsymbol{\theta}_1}(g_R, g_B)\mathcal{F}^{(\mathbf{p})}_{\boldsymbol{\theta}_2}(I^l_C(\mathbf{p})) \quad (5)$$

$\mathcal{G}_{\boldsymbol{\theta}_1}$ is a one-layer neural network learning parameters $\boldsymbol{\theta}_1 = (\theta_{11}, \theta_{12}, \theta_{13})$ with a sigmoid activation h,

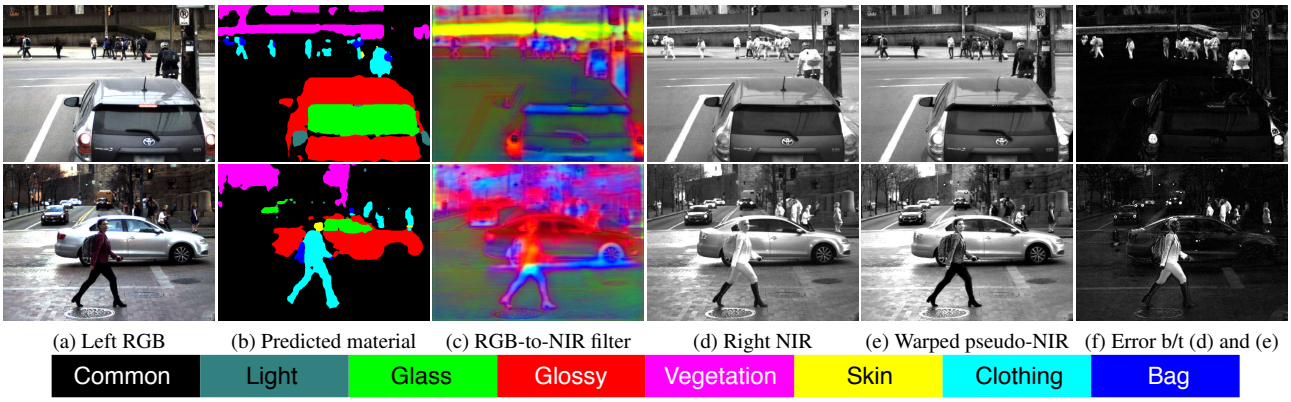
$$\mathcal{G}_{\boldsymbol{\theta}_1}(g_R, g_B) = \beta h\left(\frac{\theta_{11}}{g_R} + \frac{\theta_{12}}{g_B} + \theta_{13}\right) \quad (6)$$

where, $\beta = 2$ is the maximum white balancing gain.

$\mathcal{F}^{(\mathbf{p})}_{\boldsymbol{\theta}_2}$ calculates a weighted sum of R,G,B channels. The weights are different for each position $\mathbf{p}$. Formally,

$$\begin{aligned}&\mathcal{F}^{(\mathbf{p})}_{\boldsymbol{\theta}_2}(I^l_C(\mathbf{p})) = \\ &\theta_{21}(\mathbf{p})I^l_R(\mathbf{p}) + \theta_{22}(\mathbf{p})I^l_G(\mathbf{p}) + \theta_{23}(\mathbf{p})I^l_B(\mathbf{p})\end{aligned} \quad (7)$$

where I^l_R, I^l_G, I^l_B are the three channels of I^l_C, and the weights $\boldsymbol{\theta}_2(\mathbf{p}) = (\theta_{21}(\mathbf{p}), \theta_{22}(\mathbf{p}), \theta_{23}(\mathbf{p}))$ are predicted by a filter generating network (FGN) [8].



(a) Left RGB (b) Predicted material (c) RGB-to-NIR filter (d) Right NIR (e) Warped pseudo-NIR (f) Error b/t (d) and (e)

Figure 3. Intermediate results. (b) is the material recognition result from DeepLab [3] (explained in Section 4.2). (c) shows the RGB-to-NIR filters corrected by exposure and white balancing. The R,G,B values represent the weights of R,G,B channels. Some clothing fails in spectral translation because the relationship between its RGB and NIR intensities is low. The structural similarity term in alignment loss (Equation 3) can partially solve this problem as long as the structure is preserved.

To prevent the STN from learning disparity, we use a CNN with left-right symmetric filtering kernels (symmetric CNN) as the FGN. Thus the FGN treats the left and right parts around each pixel equally and does not shift the input and therefore learns no disparity. The structure of the FGN is the same as the DPN. The FGN accepts a RGB image and predicts a RGB-to-NIR filter (Figure 3 (c)). Yeh *et al*. [41] also proposed a symmetric filter CNN for recognition but their filters are radial symmetric while ours are reflection symmetric.

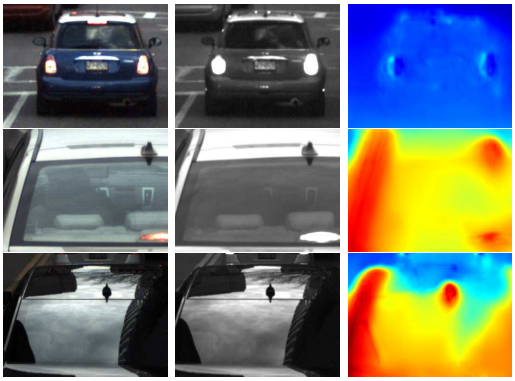
The STN loss matches the NIR and pseudo-NIR images:

$$L_{STN} = \frac{1}{N} \sum_{\mathbf{p} \in \Omega} (|I^l_{pN}(\mathbf{p}) - \tilde{I}^l_N(\mathbf{p})| + |I^r_N(\mathbf{p}) - \tilde{I}^r_{pN}(\mathbf{p})|) \tag{8}$$

where I^l_{pN}, $\tilde{I}^l_N$, I^r_N and $\tilde{I}^r_{pN}$ are, respectively, the pseudo-NIR image, the warped NIR image, the NIR image, and the warped pseudo-NIR image as defined in Section 3.1.

4. Incorporating Material-aware Confidence into Disparity Prediction Network

Though the DPN and STN work well in many cases, they cannot handle certain materials including lights, glass and glossy surfaces due to unreliable matching. Matching on these materials is hard due to large spectral change (Figure 1) and not trustworthy because it does not represent the correct disparity (Figure 4). Such materials are common but difficult to identify without external knowledge. Assessing reliability by matching score or view consistency [32, 43] fails because unreliable regions may show high matching scores (Figure 4) and strong view consistency. A light source may show a different size in RGB and NIR and thus match at its edge instead of the center. Transmitted or reflected scenes may match perfectly but the predicted disparities do not correspond to the physical surfaces.



(a) RGB patch (b) NIR patch (c) Wrong disparity

Figure 4. Unreliable matching with high matching score. (c) is predicted by DPN without material awareness. Row 1: the light sources showing different sizes in RGB and NIR, and incorrectly match at the edges instead of the centers. Row 2: matching of transmitted scene does not represent the correct windshield disparity. Row 3: disparity of the reflected scene does not correspond to the car surface.

Our goal is to incorporate material-aware confidence into DPN loss (Equation 1) to solve this problem. We propose two novel techniques: (1) Propagate the disparity from the reliable to the unreliable regions using a new confidence-weighted smoothing technique (Section 4.1) and (2) Extend the DPN loss function to be material-aware by creating material-specific alignment and smoothness losses (Section 4.2). Section 4.3 discusses how to combine those two techniques to solve specific unreliable materials.

4.1. Confidence-weighted Disparity Smoothing

Smoothing is a common technique to infer disparity in unreliable regions. However, a smoothness loss allows unreliable regions to mislead the reliable parts by forcing them to share similar disparity. As shown in Figure 5 (c), this re-

sults in the disparity at the side of the car to be misled by the wrong prediction on the windshield.

Confidence-weighted disparity smoothing uses confident disparities to "supervise" non-confident ones. Instead of fine-tuning [38] or bootstrapping [43], we change the back-propagation behavior of the smoothness loss so that it can be embedded in the DPN loss (Equation 12). Consider two neighbor pixels $\mathbf{p}_1$ and $\mathbf{p}_2$ with predicted disparities d_1 and d_2. A $L1$ smoothness loss is $L = |d_1 - d_2|$. Let $\mathbf{W}$ be all the parameters in the DPN, then $\frac{\partial L}{\partial \mathbf{W}} = \frac{\partial L}{\partial d_1}\frac{\partial d_1}{\partial \mathbf{W}} + \frac{\partial L}{\partial d_2}\frac{\partial d_2}{\partial \mathbf{W}}$. Assume that $\mathbf{p}_1$ is confident while $\mathbf{p}_2$ is unreliable. We want d_2 to follow d_1 without harming d_1. Let $\chi(\cdot)$ be the stopping gradient operator (a.k.a. 'detach' in PyTorch [33]) that acts as an identity mapping in the forward pass but stops gradients from being backpropagated through it in the backward pass. A confidence-aware loss is $L = |\chi(d_1) - d_2|$, preventing gradients being backpropagated through d_1. $\frac{\partial L}{\partial d_1}$ is set to be zero during backpropagation, *i.e.*, $\frac{\partial L}{\partial \mathbf{W}} = \frac{\partial L}{\partial d_2}\frac{\partial d_2}{\partial \mathbf{W}}$. This can be extended into a "soft" version. Generally, let $\mathbf{p}_1$ and $\mathbf{p}_2$ have confidences c_1 and c_2. We define relative confidences as $r_1 = \frac{c_1}{c_1+c_2}$ and $r_2 = 1 - r_1$, and the confidence-weighted loss as $L = r_1|\chi(d_1) - d_2| + r_2|d_1 - \chi(d_2)|$.

In practice, we consider a disparity map $d(x, y)$ and its known confidence $c(x, y)$ (defined in Section 4.3 using material). We present detailed expressions for the confidences by defining pixel neighborhood in x and y directions. The relative confidences r^+ and r^- in $x-$direction are:

$$r^+(x,y) = \chi\left(\frac{c(x+1,y)}{c(x+1,y)+c(x-1,y)}\right) \tag{9}$$

and $r^- = 1 - r^+$, where the $\chi(\cdot)$ prevents gradients to be backpropagted to the confidence. The confidence-weighted $L1$ smoothness loss along $x-$direction is:

$$\begin{aligned} L_x(d,c)(x,y) =& r^+(x,y)\left|\frac{\chi(d(x+1,y)) - d(x-1,y)}{2}\right| \\ &+ r^-(x,y)\left|\frac{d(x+1,y) - \chi(d(x-1,y))}{2}\right| \end{aligned} \tag{10}$$

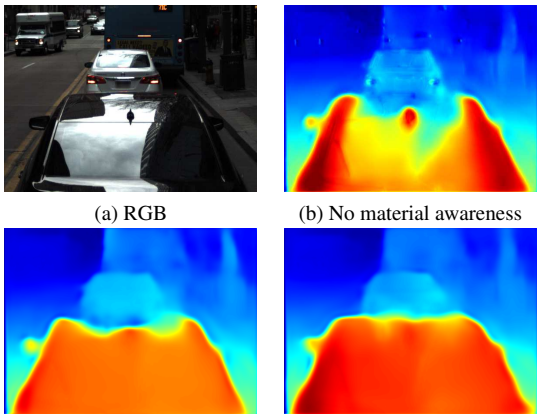
$L_y(d, c)$ is defined similarly for the $y-$direction. Then the complete confidence-weighted smoothness loss is:

$$L_{cs}(d,c) = L_x(d,c) + L_y(d,c) \tag{11}$$

As shown in Figure 5, the use of the confidence-weighted loss leads to better results than traditional smoothing.

4.2. Material-aware Loss Function

A DeepLab [3] network is used to identify unreliable regions. It is trained separately and before the training of the DPN and STN networks. A set of 8 material classes



(a) RGB (b) No material awareness (c) Smoothing w/o confidence (d) Smoothing w/ confidence

Figure 5. Comparison of smoothing with and without confidence. Smoothing without confidence makes the reliable matching around the car sides be misled by the unreliable matching on glass, which causes the predicted disparity (orange) to be smaller than the correct one (red). Introducing confidence addresses this issue.

$\mathcal{M} = \{$'light', 'glass', 'glossy', 'vegetation', 'skin', 'clothing', 'bag', 'common'$\}$ are predicted (Figure 3). 'Common' refers to any material not in the other classes. Let $\mathcal{M}^U$ be the subset of unreliable materials in $\mathcal{M}$. The DeepLab network takes a stereo pair as input and predicts left-right probabilities $\{\mu^l_m(\mathbf{p}), \mu^r_m(\mathbf{p})\}$ of each pixel $\mathbf{p}$ being material m.

To make the original DPN loss in Equation 1 material-aware, we introduce material-specific alignment and smoothness losses $L^l_{a,m}$ and $L^l_{s,m}$ respectively (similarly for the right terms). Thus, we re-write Equation 1 as:

$$\begin{aligned} L_{DPN} =& \lambda_v(L^l_v + L^r_v) \\ &+ \sum_{m\in\mathcal{M}} \lambda_{a,m}\left(\frac{1}{N}\sum_{\mathbf{p}\in\Omega}(\mu^l_m(\mathbf{p})L^l_{a,m}(\mathbf{p}) + \mu^r_m(\mathbf{p})L^r_{a,m}(\mathbf{p}))\right) \\ &+ \sum_{m\in\mathcal{M}} \lambda_{s,m}\left(\frac{1}{N}\sum_{\mathbf{p}\in\Omega}(\mu^l_m(\mathbf{p})L^l_{s,m}(\mathbf{p}) + \mu^r_m(\mathbf{p})L^r_{s,m}(\mathbf{p}))\right) \end{aligned} \tag{12}$$

For the reliable materials we use the same alignment and smoothness terms as in Equation 3 and 4, where the definition of confidence c is not required. For the unreliable materials, we use the confidence-weighted smoothness loss proposed in Section 4.1. We next describe how μ^l_m and μ^r_m are used to compute the confidence c in Equation 11.

4.3. Example Loss Terms of Unreliable Materials

Here we define the unreliable materials $\mathcal{M}^U = \{$'light', 'glass', 'glossy'$\}$ and present their loss terms.

Light Sources: Light sources like tail-lights, brake lights, bus route indicators and headlights result in unreliable matching. Thus the alignment term is $L^l_{a,light} = 0$. We assume that the light source shares the same disparity with

Method	Common	Light	Glass	Glossy	Vegetation	Skin	Clothing	Bag	Mean	Time (s)
CMA [4]	1.60	5.17	2.55	3.86	4.42	3.39	6.42	4.63	4.00	227
ANCC [17]	1.36	2.43	2.27	2.41	4.82	2.32	2.85	2.57	2.63	119
DASC [25]	0.82	1.24	1.50	1.82	1.09	1.59	**0.80**	1.33	1.28	44.7
Proposed	**0.53**	**0.69**	**0.65**	**0.70**	**0.72**	**1.15**	1.15	**0.80**	**0.80**	**0.0152**

Table 1. Quantitative results. Disparity RMSE in pixels is reported for each material. CMA [4] with searching step 0.01, ANCC [17] and DASC [25] with guided filtering [15] are tested on an Intel Core i7 6700HQ CPU. The proposed method is tested on a single NVIDIA TITAN X (Pascal) GPU. Our method outperforms the others and reaches real-time speed.

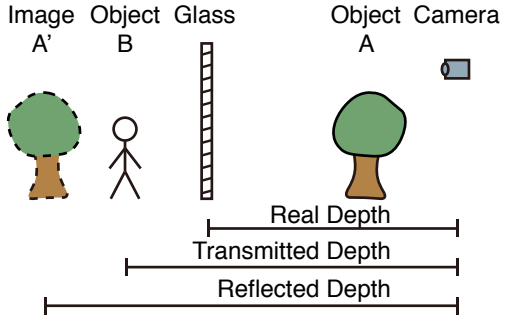


Figure 6. Transmitted and reflected scenes look farther than the real glass position.

non-light neighbors. The confidence c^l is computed using $1-\mu^l_{light}$. Then Equation 11 (smoothness term) becomes:

$$L^l_{s,light} = L_{cs}(d^l, 1-\mu^l_{light}) \tag{13}$$

Glass: Glass surfaces reflect and transmit light. We define the alignment loss $L^l_{a,glass} = 0$ considering its unreliable matching. But the dominated alignment term of common materials still forces DPN to match the appearance on glass. As illustrated in Figure 6, both the reflected and transmitted scenes appear farther than the real position of glass. Therefore, we assign higher confidence to closer scenes with larger disparities. Assuming that glass can only be physically supported by 'common', 'glass', and 'glossy' materials, we define the confidence $c^l = (\mu^l_{common} + \mu^l_{glass} + \mu^l_{glossy})e^{\frac{d^l}{\sigma}}$. Thus the smoothness loss $L^l_{s,glass}$ is:

$$L^l_{s,glass} = L_{cs}(d^l, (\mu^l_{common}+\mu^l_{glass}+\mu^l_{glossy})e^{\frac{d^l}{\sigma}}) \tag{14}$$

where, σ is a constant parameter (details in Section 6).

Glossy Surfaces: Glossy surfaces exhibit complex specular reflection. We adopt the alignment term of common materials (Equation 3), considering that it still contains some reliable matching, and the smoothness term of glass (Equation 14), because the reflected scene has smaller disparity.

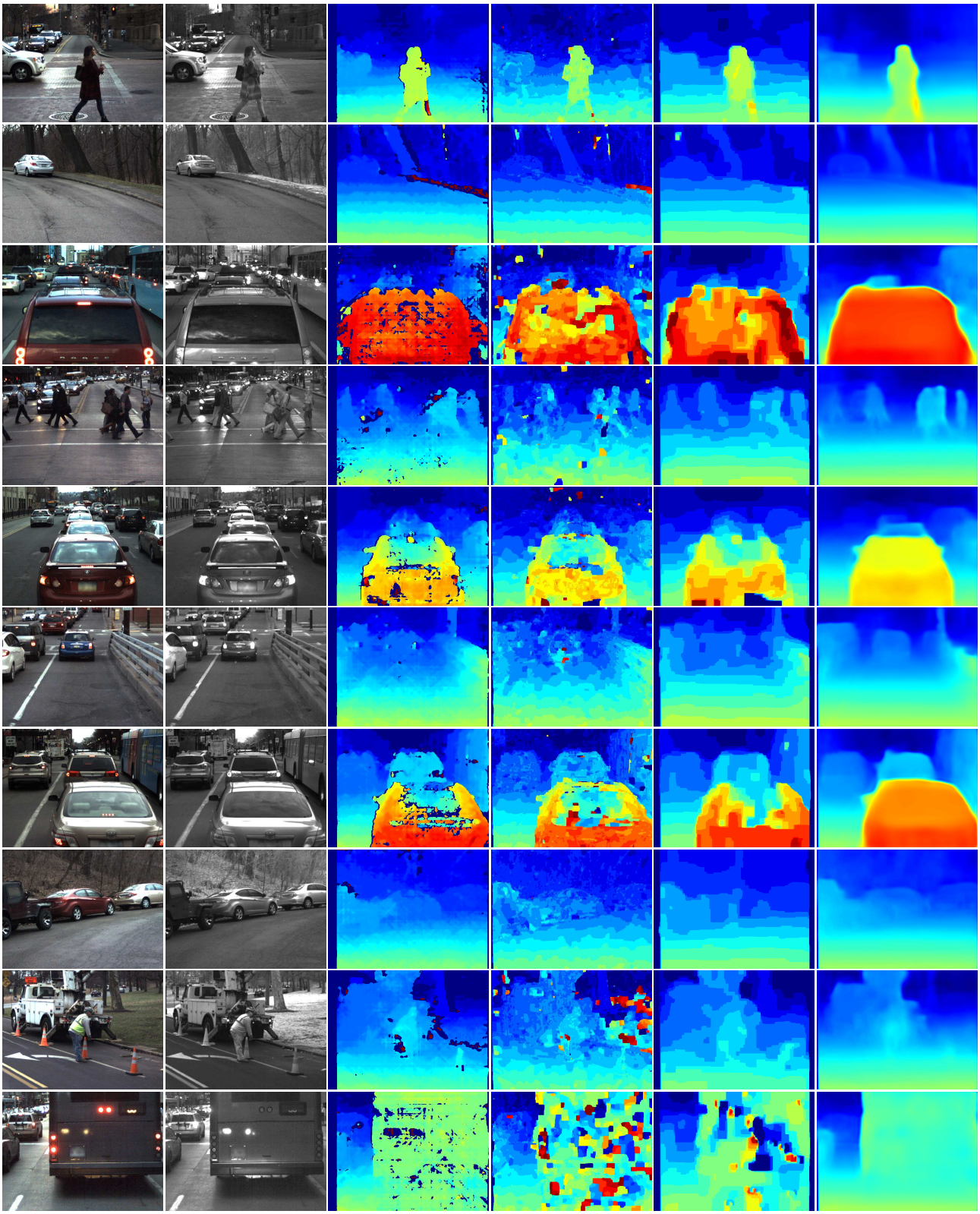
5. RGB-NIR Stereo Dataset

The dataset was captured by a RGB camera and a NIR camera mounted with $56mm$ baseline on a vehicle, alternating among short, middle and long exposures adapted by an auto-exposure algorithm at 20Hz. Close to 1 million 1164×858 rectified stereo frames equally distributed amongst the three exposure levels were collected. They were split into 12 videos, with total length of 13.7 hours. The dataset covers campus roads, highways, downtown, parks and residential areas captured under sunny, overcast and dark conditions and includes materials such as lights, glass, glossy surfaces, vegetation, skin, clothing and bags. Reliable GPS and vehicle states (speed, vehicle pose, steering radius and traveled distance) are available for 70% of the data. Images are resized into 582×429 in all experiments.

Material and disparity labels are added to a subset of the middle-exposure images. The videos are split into two sets for training (8 videos) and testing (4 videos). 3600 frames are labeled with material segments in 8 classes (common, light, glass, glossy, vegetation, skin, clothing, bag). 5030 sparse points on 2000 testing images across the 8 materials are annotated with disparity. Depth sensors are not used because they often fail on glass and light sources.

6. Experimental Results

Parameters: DPN predicts the ratio between disparity and image width. The scaling factor η is 0.008 for the DPN and 1/3 for the STN. The view consistency and alignment weights are $\lambda_v = 2$ and $\lambda_a = 1$ for all materials. The smoothness weights λ_s are 3000, 1000, and 80 for lights, glass and glossy surfaces, and 25 for other materials. The parameter in glass and glossy smoothness loss is $\sigma = 0.005$.
Training and Testing: The DeepLab [3] net is fine-tuned from a model pre-trained on ImageNet [9], COCO [30] and Pascal VOC [10]. DPN and STN are trained on $40,000$ sampled middle-exposure images with Adam optimizer [27] (batch size=16, learning rate=0.00005). They are trained with material awareness for at least 12 epochs after 4 warmup epochs without it, taking about 18 hours on two TITAN X GPUs with PyTorch [33] implementation. Only the DPN is required for testing. Negative disparities are clamped to zero.
Comparison: We have compared with Cross-Modal Adaptation (CMA) [4], ANCC [17] and DASC [25]. SIFT flow [31] search is constrained by epipolar geometry to obtain whole image disparity in DASC. Disparity RMSE (Table 1), execution times (Table 1) and qualitative results (Figure 7) are presented. Our method outperforms the others, especially on lights, glass and glossy surfaces. Our method also provides cleaner disparity maps and clearer object contours.

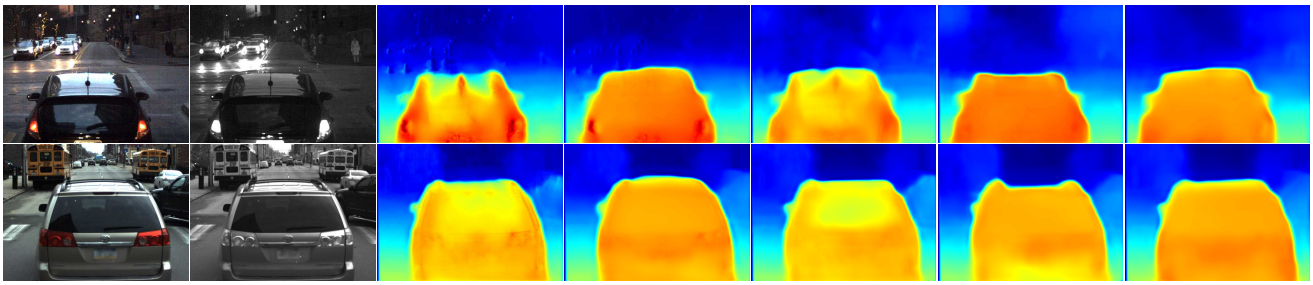


(a) Left RGB (b) Right NIR (c) CMA [4] (d) ANCC [17] (e) DASC [25] (f) Proposed

Figure 7. Qualitative results on our dataset. Image contrast is adjusted for visualization. The proposed method provides less noisy disparity maps and performs better on lights (row 3, 5, 6, 7, 10), glass (row 3, 5, 7) and glossy surfaces (row 5, 7, 10).

Method	Common	Light	Glass	Glossy	Vegetation	Skin	Clothing	Bag	Mean
Only RGB as DPN input	0.66	1.12	0.89	1.10	0.92	1.61	1.24	0.95	1.06
Averaging RGB as STN	0.52	0.80	0.74	0.78	0.76	1.30	1.21	1.04	0.89
Asymmetric CNN in STN	0.53	0.88	0.82	0.88	0.77	1.13	1.17	0.94	0.89
No material awareness	**0.51**	1.08	1.05	1.57	**0.69**	**1.01**	1.22	0.90	1.00
Ignore light sources	0.54	0.81	0.74	0.71	0.76	1.37	1.17	1.10	0.90
Ignore glass	0.56	0.74	0.97	1.08	0.75	1.06	**1.02**	0.86	0.88
Ignore glossy surfaces	0.63	0.71	0.71	1.23	0.79	1.12	1.09	0.94	0.90
Smoothing w/o confidence	0.53	**0.69**	0.71	1.20	0.85	1.06	1.12	0.81	0.87
Full proposed method	0.53	**0.69**	**0.65**	**0.70**	0.72	1.15	1.15	**0.80**	**0.80**

Table 2. Ablation study. Network structure changes (row 1-3) result in the increase of error generally. Removing material awareness (row 4-7) leads to failure on corresponding materials. Smoothing without confidence (row 8) results in performance drop. There are small fluctuations but the full method performs better in general.



(a) Left RGB (b) Right NIR (c) No material (d) Ignore lights (e) Ignore glass (f) Ignore glossy (g) Full method

Figure 8. Qualitative material ablation study. Ignoring lights results in artifacts at light sources. Ignoring glass leads to wrong disparity predictions at windshields. Ignoring glossy surfaces causes failure at the specular top surfaces of cars.

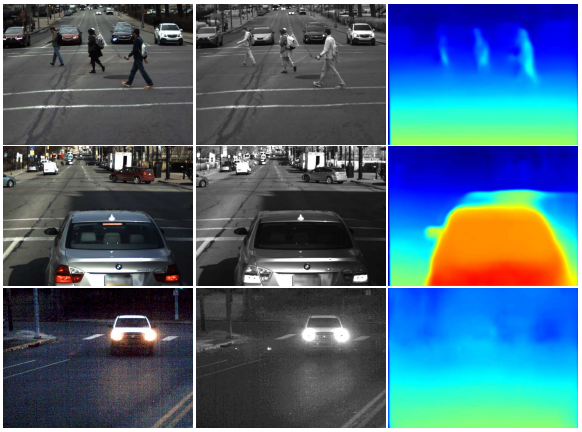
DASC performs better on clothing, possibly due to the weak relationship between its RGB and NIR appearances. Additionally, our real-time method is much faster than the others.

Ablation Study: We have tested three network structure choices: "Only RGB as DPN input", "Averaging RGB as STN" averaging R, G and B channels as pseudo-NIR, and "Asymmetric CNN in STN". Table 2 shows that overall the full method outperforms the other choices. We have also studied fully or partially removing material awareness. Table 2 and Figure 8 show that ignoring lights, glass or glossy surfaces fails on corresponding materials with small fluctuations on other materials. It means that the proposed material-specific loss functions as designed. Table 2 also shows that smoothing with confidence is useful.

7. Conclusion and Discussion

We presented a deep learning based cross-spectral stereo matching method without depth supervision. The proposed method simultaneously predicts disparity and translates a RGB image to a NIR image. A symmetric CNN is utilized to separate geometric and spectral differences. Material-awareness and confidence-weighted smoothness are introduced to handle problems caused by lights, glass and glossy surfaces. We build a large RGB-NIR stereo dataset with challenging cases for evaluation.

Our method outperforms compared methods, especially on challenging materials, although it fails on some clothing with large spectral difference, shadow edges and dark noisy regions (Figure 9). Redesigning the loss function might help address those problems. In the future, we will extend our work to other spectra (SWIR, MWIR, thermal) and to data obtained from mobile consumer devices.



(a) Left RGB (b) Right NIR (c) Predicted disparity

Figure 9. Failure cases. Row 1-3: failing to handle large spectral difference of clothing, treating shadow edge as object edge, and mismatching noise.

Acknowledgements. This work was supported in parts by ChemImage Corporation, an ONR award N00014-15-1-2358, an NSF award CNS-1446601, and a University Transportation Center T-SET grant.

References

[1] T. Bourlai, A. Ross, C. Chen, and L. Hornak. A study on using mid-wave infrared images for face recognition. In *SPIE DSS*, 2012. 1

[2] M. Brown and S. Süsstrunk. Multi-spectral sift for scene category recognition. In *CVPR*, 2011. 1

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 4, 5, 6

[4] W. W.-C. Chiu, U. Blanke, and M. Fritz. Improving the kinect by cross-modal stereo. In *BMVC*, 2011. 2, 6, 7

[5] G. Choe, S.-H. Kim, S. Im, J.-Y. Lee, S. G. Narasimhan, and I. S. Kweon. Ranus: Rgb and nir urban scene dataset for deep scene parsing. *IEEE Robotics and Automation Letters*, 2018. 1

[6] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *ICLR*, 2016. 3

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[8] B. De Brabandere, X. Jia, T. Tuytelaars, and L. Van Gool. Dynamic filter networks. In *NIPS*, 2016. 3

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 6

[11] C. Feng, S. Zhuo, X. Zhang, L. Shen, and S. Susstrunk. Near-infrared guided color image dehazing. In *ICIP*, 2013. 1

[12] R. Garg, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016. 2

[13] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 2, 3

[14] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *ECCV*, 2014. 1

[15] K. He, J. Sun, and X. Tang. Guided image filtering. In *ECCV*, 2010. 6

[16] R. He, X. Wu, Z. Sun, and T. Tan. Learning invariant deep representation for nir-vis face recognition. In *AAAI*, 2017. 1

[17] Y. S. Heo, K. M. Lee, and S. U. Lee. Robust stereo matching using adaptive normalized cross-correlation. *TPAMI*, 2011. 2, 6, 7

[18] Y. S. Heo, K. M. Lee, and S. U. Lee. Joint depth map and color consistency estimation for stereo images with different illuminations and cameras. *TPAMI*, 2013. 2

[19] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*, 2015. 1

[20] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 3

[21] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *NIPS*, 2015. 2

[22] H.-G. Jeon, J.-Y. Lee, S. Im, H. Ha, and I. So Kweon. Stereo matching with color and monochrome cameras in low-light conditions. In *CVPR*, 2016. 2

[23] N. D. Kalka, T. Bourlai, B. Cukic, and L. Hornak. Cross-spectral face recognition in heterogeneous environments: A case study on matching visible to short-wave infrared imagery. In *IJCB*, 2011. 1

[24] P. Kaur, K. J. Dana, and G. Oana Cula. From photography to microbiology: Eigenbiome models for skin appearance. In *CVPR Workshops*, 2015. 1

[25] 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, 6, 7

[26] S. Kim, D. Min, S. Lin, and K. Sohn. Deep self-correlation descriptor for dense cross-modal correspondence. In *ECCV*, 2016. 2

[27] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[28] Y. Kuznietsov, J. Stückler, and B. Leibe. Semi-supervised deep learning for monocular depth map prediction. In *CVPR*, 2017. 2

[29] J. Lezama, Q. Qiu, and G. Sapiro. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. In *CVPR*, 2017. 1

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6

[31] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *TPAMI*, 2011. 6

[32] C. Mostegel, M. Rumpler, F. Fraundorfer, and H. Bischof. Using self-contradiction to learn confidence measures in stereo vision. In *CVPR*, 2016. 4

[33] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS Workshops*, 2017. 5, 6

[34] P. Pinggera, T. P. Breckon, and H. Bischof. On cross-spectral stereo matching using dense gradient features. In *BMVC*, 2012. 2

[35] D. Rüfenacht, C. Fredembach, and S. Süsstrunk. Automatic and accurate shadow detection using near-infrared information. *TPAMI*, 2014. 1

[36] X. Shen, L. Xu, Q. Zhang, and J. Jia. Multi-modal and multi-spectral registration for natural images. In *ECCV*, 2014. 2

[37] S. Song and J. Xiao. Tracking revisited using rgbd camera: Unified benchmark and baselines. In *ICCV*, 2013. 1

[38] A. Tonioni, M. Poggi, S. Mattoccia, and L. Di Stefano. Unsupervised adaptation for deep stereo. In *ICCV*, 2017. 2, 5

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 3

[40] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe. Learning cross-modal deep representations for robust pedestrian detection. In *CVPR*, 2017. 1

[41] R. Yeh, M. Hasegawa-Johnson, and M. N. Do. Stable and symmetric filter convolutional neural network. In *ICASSP*, 2016. 4

[42] X. Zhang, T. Sim, and X. Miao. Enhancing photographs with near infra-red images. In *CVPR*, 2008. 1

[43] C. Zhou, H. Zhang, X. Shen, and J. Jia. Unsupervised learning of stereo matching. In *ICCV*, 2017. 2, 4, 5

[44] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 2