IJCV manuscript No.
(will be inserted by the editor)

CMSNet: Deep Color and Monochrome Stereo

Hae-Gon Jeon[1] · **Sunghoon Im**[2] · **Jaesung Choe**[3] · **Minjun Kang**[3] · **Joon-Young Lee**[4] · **Martial Hebert**[5]

Received: date / Accepted: date

Abstract In this paper, we propose an end-to-end convolutional neural network (CNN) for stereo matching with color and monochrome cameras, *called* CMSNet (Color and Monochrome Stereo Network). Both cameras have the same structure except for the presence of a Bayer filter, but have a fundamental trade-off. The Bayer filter allows capturing chrominance information of scenes, but limits a quantum efficiency of cameras, which causes severe image noise. It seems ideal if we can take advantage of both the cameras so that we obtain noise-free images with their corresponding disparity maps. However, image luminance recorded from a color camera is not consistent with that from a monochrome camera due to spatially-varying illumination and different spectral sensitivities of the cameras. This degrades the performance of stereo matching. To solve this problem, we design CMSNet for disparity estimation from noisy color and relatively clean monochrome images. CMSNet also infers a noise-free image with the estimated disparity map. We leverage a data augmentation to simulate realistic signal-dependent noise and various radiometric distortions between input stereo pairs to train CMSNet effectively. CMSNet is evaluated using various datasets and the performance of our disparity estimation and image enhancement consistently outperforms state-of-the-art methods.

Keywords Stereo matching · disparity estimation · image enhancement · convolutional neural network

Hae-Gon Jeon
E-mail: haegonj@gist.ac.kr

Sunghoon Im
E-mail: sunghoonim@dgist.ac.kr

Jaesung Choe
E-mail: jaesung.choe@kaist.ac.kr

Minjun Kang
E-mail: kmmj2005@kaist.ac.kr

Joon-Young Lee
E-mail: jolee@adobe.com

Martial Hebert
E-mail: hebert@ri.cmu.edu

[1] AI Graduate School & The School of Electrical Engineering and Computer Science, GIST, Gwangju, Korea, [2] The Information and Communication Department, DGIST, Deagu, Korea, [3] KAIST, Daejeon, Korea, [4] Adobe Research, San Jose, CA, USA, [5] The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA

1 Introduction

As the computing power of hand-held devices grows, multi-camera setups (Hua 2016; V50 2019; Gal 2019; iPh 2018) have become available and those allow us to acquire scene depths from simultaneously captured images. Interestingly, the commercial products have asymmetric camera configurations (Shen et al 2017) which have different field-of-views (FOVs) (V50 2019; Gal 2019; iPh 2018) or different spectral properties (Hua 2016). Among them, the RGB-Monochrome setup has been commercialized, e.g., Huawei P9, P10, and P20 series (Hua 2016). In this system, a color camera has a Bayer color filter in front of an image sensor to separate the incoming light into one of three primary colors (red, green, or blue) by filtering the light spectra according to corresponding wavelength ranges. Although this process is effective to capture color information, it amplifies image noise in low-light conditions because the array occludes a lot of incoming light. Unlike color cameras, monochrome cameras receive all the incoming light at each pixel and need no demosaicing process. Therefore, those have much better light efficiency and pro-

vide sharper images than Bayer-filtered color cameras within the same spectral band as illustrated in Fig. 1.

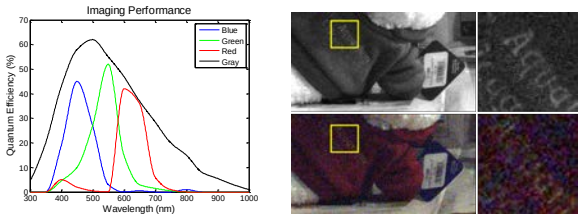
In this paper, we present a stereo matching and colorization network that takes advantage of both color and monochrome cameras. Our CMSNet achieves the color sensing capability on color cameras and light efficiency on monochrome cameras. However, a non-linear spectral sensitivity between the asymmetric image pairs makes it hard to find accurate correspondences (Jeon et al 2016). To tackle this issue, we leverage convolutional neural network (CNN) for stereo matching and high-quality color image recovery from a single Color-Monochrome image pair. We call this end-to-end network as CMSNet. Our CMSNet consists of four sub-networks: (1) disparity estimation, (2) RGB and monochrome image denoising, (3) occlusion detection, and (4) colorization. We first estimate a disparity map from an RGB-Monochrome image pair, and denoise each input image. With the estimated disparity map, we make an initial recovered color image by transferring chrominance of the denoised color image into the denoised monochrome image. The initial recovered color image suffers from color bleeding errors due to occluded regions between stereo images or inaccurate disparities. In order to correct the errors, CMSNet infers an occlusion map from the input image pair, then it produces a final high-quality color image using the colorization network.

This paper is an extended version of our previous work that devises an iterative stereo matching and propagates chrominance channels of an input color image with the estimated disparity map to an input monochrome image. Inspired by (Jeon et al 2016), CMSNet takes an end-to-end deep network which allows us to simultaneously infer a disparity map and a high-quality color image within short inference time. Additionally, this paper provides more in-depth analysis and experimental results to validate the effectiveness of CMSNet.

The reminder of the paper is organized as follows. In Section 2, we review state-of-the-art low-light imaging and cross-spectral/multi-modal stereo matching methods. In Section 3, we describe the advantages of our proposed color and monochrome stereo system. We present CMSNet including its analysis and implementation details in Section 4. We then show the robustness of our approach compared to the other algorithms in Section 5. Finally, we draw conclusions in Section 6.

2 Related Work

This work is related to low-light photography, cross-spectral stereo matching, and colorization. Prior to introducing previous studies, we refer the reader to the



(a) (b)

Fig. 1 (a) Spectral sensitivity of the color and monochrome camera (Fle 2017) used in our prototype stereo system. (b) An example image pair captured by our stereo system. Note that there is visible difference of image noise due to the gap of light-efficiency between two cameras.

work (Hirschmüller and Scharstein 2009) for the comprehensive discussion of stereo matching with radiometric and noise variations.

2.1 Low-light Photography

There are various ways to take high-quality photos in low-light environments. The most straightforward way is a single image denoising (Dabov et al 2007; Buades et al 2005), but it often suffers from over-smoothing artifacts. The over-smoothing artifacts come mostly from a relatively large window of neighboring pixels to suppress severe noise. Single image-based approaches have been recently boosted up via CNN architectures (Zhang et al 2017a; 2018).

As an alternative way, multiple image averaging methods have gained research interests (Liu et al 2014; Hasinoff et al 2016). Well-aligned images are beneficial for image denoising because they take an align-and-average strategy in usual. Using a burst mode of off-the-shelf cameras, sequential images are captured with short and consistent exposure times, and then those are merged by a local homography (Liu et al 2014) and a sub-pixel alignment in frequency domain (Hasinoff et al 2016). Even though short and consistent exposure allows the methods to easily find correspondences between input images, such images may include many under-exposed regions.

High dynamic range (HDR) (Reinhard et al 2010) and exposure fusion (Mertens et al 2009) have been considered as the most representative solution to the under-exposure problem. In addition, these recent methods address the challenges of varying exposure with sophisticated alignment and inpainting (Gallo et al 2009; Hu et al 2013). However, the performance is not guaranteed for datasets with fast non-rigid parts of scenes which cause inaccurate merging results, even with the advanced CNN-based alignment in (Im et al 2019a).

2.2 Cross-channel and Multi-spectral Stereo

Cross-spectral stereo matching has been extensively studied to find correspondences of either multi-modal or color-inconsistent stereo images. Heo *et al.* (Heo et al 2011) analyzed a color formation model and proposed an adaptive normalized cross correlation for stereo matching, which is robust to various radiometric changes. It is extended in (Heo et al 2013), which presented an iterative framework to simultaneously achieve both depth estimation and color consistency. Pinggera *et al.* (Pinggera et al 2012) presented depth map estimation with cross-spectral stereo images, which uses dense gradient features based on the HOG descriptor (Dalal and Triggs 2005). Kim *et al.* (Kim et al 2015) designed a dense descriptor for multi-modal correspondences by leveraging an adaptive self-correlation measure and a randomized receptive field pooling. Holloway *et al.* (Holloway et al 2015) proposed an assorted camera array and a cross-channel point correspondence measure using a normalized gradient cost. In (Zhi et al 2018; Liang et al 2019), unsupervised CNN architectures are proposed for stereo matching with RGB-NIR image pairs. Both the methods devise their own spectral translation networks to convert an RGB image into a pseudo-NIR image and left-right consistencies based on the estimated disparity map to handle the matching problem in large appearance changes in different spectra.

Compared to the previous studies, we focus on simultaneously reconstructing an accurate disparity map and a noise-free color image with a color and monochrome image pair. We achieve this by taking the advantage of our cross-spectral stereo system. We present a CNN architecture with depth supervision which consists of three encoder-decoder structures for disparity estimation, occlusion detection, and colorization-based color image recovery as well as one denoising network. As will be demonstrated in the experiment section in Section 5, our method is highly effective for accurate disparity estimation and largely outperforms the state-of-the-art algorithms (Heo et al 2011; 2013; Kim et al 2015; Holloway et al 2015; Zhi et al 2018). In colorization, most approaches concentrate on propagating limited numbers of user-defined seeds, while we have lots of seed pixels with outliers around occlusion boundaries. To handle this issue, we introduce an occlusion map to correct inaccurate seed pixels and successfully recover a high-quality color image.

2.3 Colorization

Colorization is a process of adding color channels to grayscale image and video. Levin *et al.* (Levin et al 2004) presented a user-guided colorization method which takes partial color information from user scribbles and automatically propagates the given seed color to make a complete color image. Yatziv and Sapiro (Yatziv and Sapiro 2006) proposed a fast colorization method using a geodesic distance between neighboring pixels. Gastal and Oliveira (Gastal and Oliveira 2011) introduced an edge-aware filter in a transformed domain and presented colorization results. Zhang *et al.* (Zhang et al 2016) showed that a CNN is able to automatically predict color values from a single grayscale image without any user interaction. In (Zhang et al 2017b), a CNN architecture learns an initial seed suggestion and propagation of the initial seed accurately. Irony *et al.* (Irony et al 2005) proposed an example-based colorization with an assumption that similarly textured regions have similar colors. A CNN version of the example-based colorization is proposed in (He et al 2018). The work computes a similarity score between reference images and unaligned target images to infer aligned chrominance channels prior to colorization.

Another related work is (Chakrabarti et al 2014), which presented the concept of an alternative camera sensor that samples color information very sparsely. They recover a full color image by propagating the sparsely sampled colors into an entire image. This work shares the same philosophy with our work that takes the advantage of light-efficient monochrome sensors but the method may suffer from color noise which leads to an erroneous color image. In (Dong et al 2019), a CNN-based colorization in color-monochrome stereo system is proposed. The CNN computes weighted average of colors of candidate pixels in reference image for initial chrominance assignments, and a color residual module to correct inaccurate assigned pixels caused by occluded areas. We adopt similar ideas to a stereo system and obtain an accurate depth-map and a noise-free color image simultaneously.

3 Color/Monochrome Stereo System

Most color cameras use a color filter array called a Bayer array to capture color information. The Bayer array is positioned over the pixels of an image sensor and separates the incoming light into one of three primary colors (red, green, or blue) by filtering the light spectra according to corresponding wavelength ranges. This process is effective to capture color information, but it amplifies image noise in low-light conditions since the array occludes a lot of incoming light. It may also reduce image sharpness by an anti-aliasing filter or optical low-pass filter to avoid aliasing or moiré artifacts during the demosaicing process (Kimmel 1999).

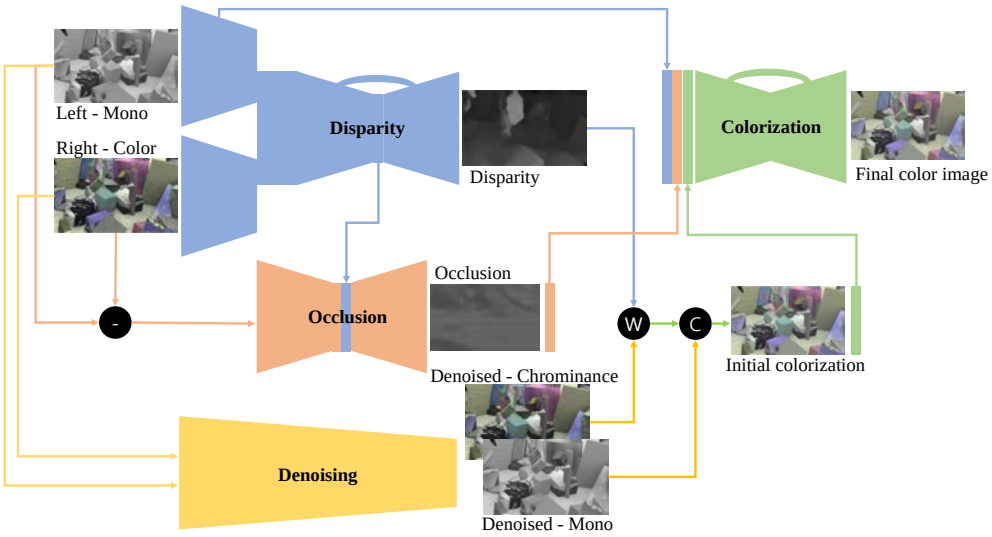


Fig. 2 Illustration of our CMSNet architecture. Given color and monochrome stereo pair, the disparity estimator (blue) yields disparity map corresponding to the left monochrome image, and the sub-network for denoising (yellow) removes image noise. With the estimated disparity map and denoised images, the colorization sub-network makes a high-quality color image. At this point, the occlusion estimator (orange) infers an occlusion map to handle color bleeding errors which is used for the colorization. **-**,**W** and **C** denote a subtraction, warping and concatenation along with a channel axis.

Unlike color cameras, monochrome cameras receive all the incoming light at each pixel and need no demosaicing process. Therefore, those have much better light efficiency and provide sharper images. In Fig. 1, we compare the imaging quality of a color and a monochrome camera. The comparison of spectral sensitivity (Fig. 1(a)) and the example image pair captured in the same condition (Fig. 1(b)) prove the large difference of light efficiency and image quality between two cameras. That is, a color and monochrome camera pair are highly suitable to achieve a noise-free color image in addition to accurate depth estimation.

4 Deep Color and Monochrome Stereo Network

We present a color and monochrome stereo network whose final goal is to simultaneously output a disparity map and a high quality color image. The proposed network consists of four sub-networks: disparity, occlusion, denoising, and colorization networks. The first three networks take the color and monochrome stereo images as input in order to compute a disparity map, an occlusion map, and recovered images, respectively. With the estimated disparity, we then warp chrominance channels of the input color image into the monochrome image viewpoint to generate an initial colorized image. The colorization network infers a high quality image from the estimated disparity map, occlusion map and the initial colorized image. The occlusion map helps to remove inaccurately assigned initial seeds of the chrominance channels. Our approach is superior to existing cross-channel and multi-spectral stereo matching methods quantitatively and qualitatively. The success comes from a proper combination of four sub-networks whose overview and detail are shown in Fig. 2 and Sec. 7, respectively.

4.1 Disparity Estimation

The overall structure of our disparity estimation network is similar to (Mayer et al 2016). The network has an encoder-decoder structure and has two input streams for each input image. We connect feature maps between corresponding layers of encoder and decoder to preserve both high-level information and fine local information.

In traditional stereo matching (Hirschmüller and Scharstein 2009), a cross correlation that performs multiplicative patch comparisons to account for gain and radiometric changes. To take the advantage of the cross correlation, its modifications are used for multi-spectral/cross-channel stereo matching (Heo et al 2011; Holloway et al 2015). Several variants of the cross correlation are designed to impose high weights for the image textures with different spectral ranges of wavelengths. Particularly, the work in (Holloway et al 2015) computes matching costs only using gradient patches of cross-/multi-spectral images because they barely share photo-consistency in practice.

We utilize it to compute matching costs between two learned features (Dosovitskiy et al 2015) as well. The size of feature maps from the correlation layer is four-dimensional: for every combination of two 2D positions we obtain a correlation value, i.e. the scalar product of the two vectors which contain the values of the cropped patches, respectively. In practice, we organize the disparity in channels, whose range is a hyper-parameter.

A concatenation of two learned feature maps could also be considered because it shows better efficiency on the cost volume-based stereo matching (Kendall et al 2017; Chang and Chen 2018; Im et al 2019b). The concatenation has an opportunity to learn an absolute representation and carry this through to the cost volume. However, in this work which directly regresses correspondences like flownet (Dosovitskiy et al 2015), the cross correlation is more effective than the concatenation since the correlation layer explicitly provides matching capabilities.

In order to optimize this sub-network, we use a berHu loss (as known as reverse Huber loss) (Owen 2007) as follow:

$$E_D = \frac{1}{N}\sum_i B(D^i_{est}, D^i_{gt}),$$
$$\text{s.t.} \quad B(x,y) = \begin{cases} |x-y| & \text{if} \quad |x-y| \le c \\ \frac{|x-y|-c^2}{2c} & \text{otherwise} \end{cases} \tag{1}$$

where D_{est} and D_{gt} are the estimated and ground-truth disparity maps, respectively. N is the number of pixels of the image and i denotes a pixel index of the image. $|\cdot|$ is an absolute operation and c is a variable assigned as $\alpha \max_i |D^i_{est} - D^i_{gt}|$ with $\alpha = 0.2$ in (Tosi et al 2019).

4.2 High-quality Color Image Recovery

In (Jeon et al 2016), the YUV colorspace is used for color image recovery by compositing one luminance channel Y of the monochrome image and two chrominance channels, U and V, of the color image. They directly use the monochrome input image as the luminance channel of a recovered color image and reconstruct color information of it by combining the chrominance channels of the color input image according to the estimated disparity. The recovered color image often suffers from incorrect chrominance mapping and color bleeding errors due to occlusion and an inaccurate disparity map. These errors are corrected with a modified colorization algorithm (Levin et al 2004) that segments the luminance channel into super-pixels (Achanta et al 2012) and computes the confidence of initial chrominance mapping. Additionally, since the colorization (Levin et al 2004) is similar to neighborhood propagating properties, it enjoys its built-in smoothing effect on the color image recovery. Motivated from it, we separate the color image recovery task as in (Jeon et al 2016).

(a) GT (b) denoised input (c) noisy input

Fig. 3 A comparison of depth maps obtained from denoised and noisy input stereo image.

4.2.1 Denoising

We first remove the noise of each input image using a denoising sub-network. The sub-network has the same structure with (Zhang et al 2017a). The network consists of 17 convolutional layers with Batch normalization and ReLU activation except for the first and last layers, and outputs a residual image for noise. The benefit of Batch normalization is that the residual output follows a Gaussian distribution which facilitates the Gaussian normalization step of batch normalization. In addition, we note that each convolution layer is initialized by orthogonal regularization, working well in suppressing image noise and preserving details.

The sub-network is optimized by minimizing a Euclidean distance between input image and denoised images as below:

$$E_N = \frac{1}{N \times CH}\sum_{ch}\sum_i \|G^i_{ch} - (I^i_{ch} - R^i_{ch})\|, \tag{2}$$

where G is a ground-truth image, I is a noisy image and R is a residual output from the sub-network. $\|\cdot\|$ denotes a L_2 norm. ch represents the index of image channels. CH denotes the number of color channels and is set to 1 if an input is a monochrome image. We note that the disparity estimation in Sec. 4.1 is performed with only noisy image pairs because a denoised stereo pair are not matchable (see Fig. 3).

4.2.2 Occlusion Estimation

Prior to colorization, we estimate occlusion regions that lead to color mapping errors even for the pixels warped by accurate disparity values (see Fig. 4(a) and (b)).

There are two types of occlusion estimation using CNNs. The first one is to estimate occlusion maps and disparity maps separately in (Li and Yuan 2018). The work infers occlusion maps from stereo image pairs at

first, and then computes disparity maps with the stereo pairs and the estimated occlusion maps. The second manner shares a common encoder to extract feature maps from stereo image pairs, and uses separate decoders to produce disparity maps and occlusion maps with their consistency term in a loss function (Zhao et al 2020).

For this, we design a separate encoder-decoder structure network. This sub-network takes a difference image between the monochrome image and the decolorized color image which is warped by the estimated disparity map as input. In addition, a compressed feature from the encoder of disparity estimation is utilized to embed implicit disparity information between them. Similar with recent cross-/multi-spectral matching (Quan et al 2019), the difference as input aims to extract a feature map for texture similarities.

The feature map is embedded with a feature map from the encoder for disparity estimation in Sec. 4.1 before passing through the decoder. This aims to utilize learned feature maps for parallax between stereo images as well as texture similarities. We observe that both feature maps make more reliable predictions.

As a loss function, we use a binary cross-entropy as below:

$$E_O = -\frac{1}{N}\sum_i O_i \log(p(O_i)) + (1 - O_i)\log(1 - p(O_i)), \tag{3}$$

where O is the binary mask (1 for occluded regions and 0 for unoccluded regions) and $p(\cdot)$ is the predicted probability of the occluded regions.

4.2.3 Colorization

Similar to (Jeon et al 2016), our colorization sub-network ingests images in the YUV colorspace which is composed of one luminance channel Y in the monochrome image and two chrominance channels, U and V in the warped color image through the estimated disparity map. The warping is performed by differentiable bilinear interpolation (Jaderberg et al 2015). We additionally use the estimated occlusion map by concatenating it with the initial colorized image to correct color bleeding errors to impose a weight for occluded regions (see Fig. 4(c) and (d)).

The sub-network aims to refine the initial colorized image by removing bleeding color pixels and by re-inferring correct chrominance values. Here, the occlusion map is used to impose a weight for occluded regions which potentially cause inaccurate colorization results. In addition, the sub-network considers the content of the monochrome image with the chrominance information during the colorization process. Since the monochrome image is a reference image, the chrominance acts as initial seeds, similar to edge-aware filters like Gastal and Oliveira (2011). The occlusion map stops spreading the chrominance information into unreliable regions.

For colorization, we use a U-Net structure (Ronneberger et al 2015), which has been shown to work well for a colorization task (Zhang et al 2017b). Our colorization sub-network has 8 convolutional blocks consisting of Conv-Conv-BatchNorm (block 1 to 4) and Deconv-Conv-Conv-UpSample (block 5 to 8). All convolutional layers are followed by ReLU activation except for the prediction layer. In the encoding layers, feature maps are progressively halved spatially, while doubling in the feature dimensions. In the decoding layers, spatial resolution is recovered, while feature dimensions are halved. Skip connections are added to help the network recover spatial information similar to the disparity estimation in Section 4.1. Our sub-network for colorization does not require any user input. In addition, it does not need to compute additional statistics because we have enough initial seeds through the color mapping. Instead of stacking all inputs together, our sub-network also concatenates the inputs after passing through each convolution layer.

The sub-network is optimized by minimizing a Euclidean distance between a ground-truth chrominance C and a recovered chrominance J as below:

$$E_C = \frac{1}{2N}\sum_{k\in\{U,V\}}\sum_i \|C_k^i - J_k^i)\|. \tag{4}$$

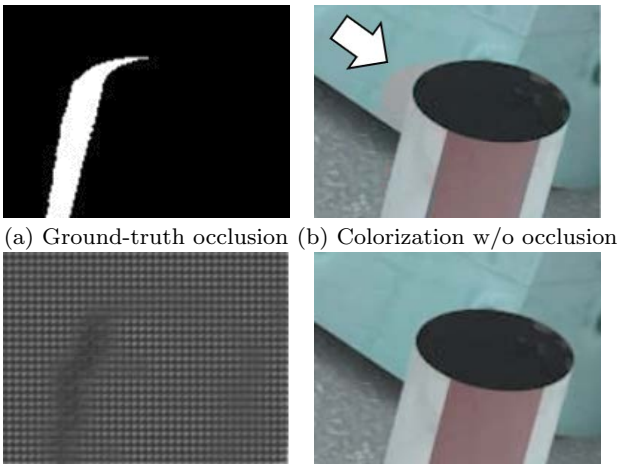
In total, CMSNet uses one loss function which is a linear combinations of Eq.1, 2, 3 and 4 as below:

$$E = E_D + \lambda_1 E_N + \lambda_2 E_O + \lambda_3 E_C, \tag{5}$$

where λ_1, λ_2 and λ_3 are hyper-parameters of our loss function. We empirically set them to 1, 100, 0.1, respectively. CMSNet dose not utilize any pre-trained models and we will describe the training details in Sec. 4.4.

4.3 Data Augmentation

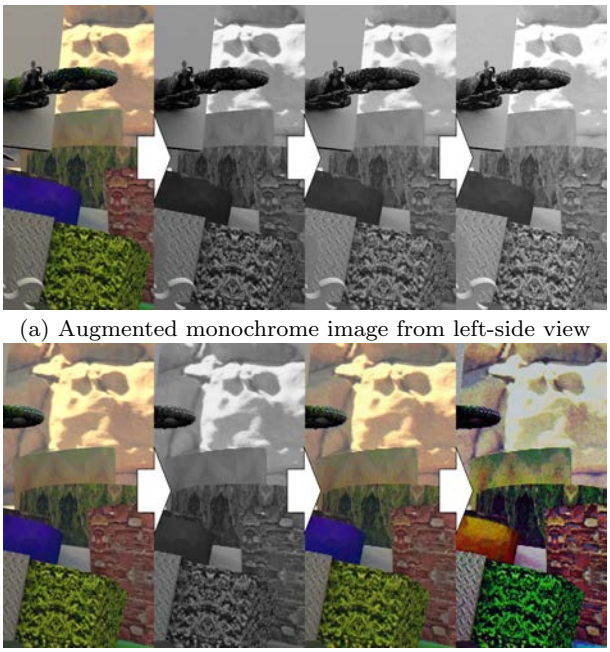
There is no public RGB-monochrome stereo image dataset. By contrast, many color stereo image pairs are available. In this environment, we have to generate RGB-monochrome image pairs for training CMSNet from public synthetic stereo datasets Mayer et al (2016). For this, we augment the synthetic stereo datasets for RGB and monochrome images separately. This augmentation is based on the properties of each camera as discussed in Sec. 3.



(a) Ground-truth occlusion (b) Colorization w/o occlusion

(c) Estimated occlusion (d) Colorization w/ occlusion

Fig. 4 An example of the effectiveness of the estimated occlusion map in colorization.



(a) Augmented monochrome image from left-side view

(b) Augmented color image from right-side view

Fig. 5 An example of data augmentations. (a) Input color image, decolorization, random gamma mapping and adding signal dependent Gaussian noise (left to right). (b) Input color image , Raw-style image generation, adding signal dependent Gaussian noise and demosaicing and random tuning of contrast, hue and saturation (left to right).

For monochrome images (left-side view), we first decolorize them using a conventional RGB to grayscale conversion. We apply a random gamma mapping [0.6, 0.9] to imitate the light efficiency of monochrome cameras, and then randomly add signal dependent Gaussian noise with a given standard deviation where $\kappa \in \{0.005, 0.015\}$ represents the noise-free signal intensity (Achanta et al 2007) (see Fig. 5 (a)).

For color images (right-side view), we consider a demosaicing process and simulate lower resolution and less light efficiency than those of monochrome images. We sample pixels of each color channel on a fixed grid with stride 2. We then collect the pixels on a rectangle grid with 2×2 pixels in order to produce a RAW format which is a grayscale with a BGGR Bayer pattern. Since the lower resolution of color images than that of monochrome images is associated with demosaicing, we reconvert the RAW image to a color image using a conventional demosaicing (Malvar et al 2004). Before adding random signal dependent noise with higher standard deviation of $\kappa \in \{0.03, 0.05\}$, we additionally use color augmentations including saturation [0, 1], contrast [0, 2] and hue [-1, 1] in order to simulate different spectral sensitivities (see Fig. 5 (b)).

We highlight that our generated RGB-Monochrome stereo dataset achieves a generality of our network, and the experimental results show the dataset is enough to train. The synthetic datasets used never suffer from the loss of image quality. With the lossless images, we convert color images into monochrome images, and downsample color images to simulate artifacts by following a conventional demosaic process. As demonstrated in Shin et al (2018), the aggressive and proper data augmentation scheme has a significant impact on the generality of CNNs, similar to our CMSNet. We note that it is a strength of our CMSNet in that it shows promising performance without any fine-tuning with real-world images whose construction is infeasible.

4.4 Training Details

In training, we use stereo image pairs with 192×320 resolutions, its corresponding ground-truth disparity maps and occlusion maps from the train split of FlyingThings3D[1] (Mayer et al 2016). We set a disparity range to [0, 45] pixels. We randomly shuffle the whole dataset and train our model with about 35M parameters from scratch for 1M iterations in total without any further fine-tuning. We use random initialization for all convolution filters. All models were trained end-to-end with the ADAM optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$). We use a batch size of 4 and set a learning rate to $1e{-}4$ for all iterations. The training is performed with TensorFlow (Abadi et al 2015) on one Nvidia 1080 Ti GPU and it takes about three days. For inference, CMSNet

[1] Downloaded from https://lmb.informatik.uni-freiburg.de/resources/datasets/SceneFlowDatasets.en.html

takes 0.04 seconds per one disparity map and recovered color image.

5 Experiments

We demonstrate the performance of CMSNet on disparity estimation and color image recovery. For evaluation, we compare CMSNet with state-of-the-art methods of cross-channel/multi-spectral stereo matching and single/ multi-image denoising. In addition, ablation studies indicate that each of these technical contributions leads to appreciable improvements in disparity accuracy and recovered image quality.

5.1 Disparity Estimation

We quantitatively evaluate CMSNet on Flyingthings3D (test split), Middlebury stereo (Hirschmüller and Scharstein 2009), Monkaa (Mayer et al 2016) and KITTI (Menze and Geiger 2015). In our evaluation, we randomly select 50 images from each dataset and report common quantitative measures of disparity quality for 200 images in total: root mean square error (RMSE) and bad pixel ratio (BPR: A percentage of disparity error over 1, 2 and 4 pixels). Since the KITTI dataset provides sparse depth measurements from a Lidar sensor, we measure the errors for the valid pixels.

For the Flyingthings3D and the Monkaa datasets, we generate color-monochrome stereo pairs with the same manner of Section 4.3. For realistic simulation using the Middlebury dataset, we take two images captured under different illuminations to simulate different spectral sensitivities and add additional noise to simulate low-light conditions. To imitate the light-efficiency difference between color and monochrome cameras, we use longer exposure images as monochrome input images and add more noise to color input images as well. To validate a generality of CMSNet over real-world data, we use the KITTI dataset which provides two types of stereo pairs captured with color and monochrome stereo setup. We take one color and one monochrome images from these four images. Since the goal of KITTI stereo evaluation is to demonstrate the generality on the real-world scenes, we do not add noise and try color variations.

For evaluation, we compare CMSNet with state-of-the-art methods of multi-spectral or cross-channel stereo matching; ANCC (Heo et al 2011), JDMCC (Heo et al 2013), DASC (Kim et al 2015), CCNG (Holloway et al 2015) and ITER (Jeon et al 2016), which are based on hand-craft matching costs. For a fair comparison, we used the original authors' code and chose the best performing parameters after parameter sweeps. We also perform a comparison with an unsupervised multi-spectral stereo matching network, namely DMC (Zhi et al 2018). Taking advantage of the unsupervised manner, we additionally report quantitative results from DMC fine-tuned on train splits of each dataset. The quantitative comparison is presented in Table 1, whose examples are shown in Fig. 6. We also report inference time of ours and the competitive methods in Table 2.

In this experiment, the modified normalized cross correlation-based methods (Kim et al 2015; Holloway et al 2015; Heo et al 2011) are vulnerable to a low intensity level and severe noise as demonstrated in (Hirschmüller and Scharstein 2009). JDMCC (Heo et al 2013) works relatively well among the competing methods, but it exhibits large quantization errors as shown in Fig. 6 (4^{th} row). We conjecture that the absence of color information occurs the failure of color equalization of JDMCC and it yields large errors in the results of JDMCC. ITER also shows relatively good results because its decolorization for noisy color images reduces image noise during iterative matching procedure. However, tree-based filtering as a post processing sometimes causes depth quantization artifacts for unmatched pixels.

DMC does not work well in these experiments. DMC computes a color consistency between an NIR image (monochrome image here) and a pseudo NIR image from their spectral conversion network as a loss function. It is reasonable only if the RGB-NIR cameras are radiometrically calibrated and their varying radiometric setting such as white balancing gains and exposure times are known. In this experiment, we perform various color augmentations for test splits of Flyingthings3D and Monkaa, which makes DMC difficult to produce reliable disparity maps. We observe that DMC produces relatively reasonable results on the KITTI dataset having consistent scene configurations and radiometric settings.

On the other hand, CMSNet largely outperforms all the competing methods over all test datasets. Compared to the hand-craft matching methods, the proposed method causes less quantization artifact while preserving the sharp object boundaries. We observe that the proposed method with the proper data augmentation technique improves the performance of correspondence search between the multi-spectral stereo images. We also can see that our disparity network produces accurate disparity maps that represent fine details. In particular, CMSNet can be boosted up by fine-tuning a specific dataset like the Middlebury dataset[2] in this experiment.

[2] We randomly crop the images to augment and generate occlusion maps by crosschecking a pair of disparity maps.

Table 1 Comparison results. **Disparity estimation**: DASC (Kim et al 2015), CCNG (Holloway et al 2015), ANCC (Heo et al 2011), JDMCC (Heo et al 2013), ITER (Jeon et al 2016), DMC (Zhi et al 2018) and ours including ablation studies. The 'RGB pair' is to use color stereo image pairs as inputs, and the 'feat. concat' is the learned feature concatenation instead of cross correlation. The 'denoise' is to take color and monochrome images passing through our denoising network as inputs. The numbers in parentheses indicate the errors in the predicted disparity map before fine-tuning the Middlebury dataset. **Best**, Second best.

Method	FlyingThings3D				Middlebury			
	RMSE	Bad1.0	Bad2.0	Bad4.0	RMSE	Bad1.0	Bad2.0	Bad4.0
Hand-craft matching cost								
ANCC (Heo et al 2011)	8.7748	0.5270	0.3640	0.2673	5.0382	0.4625	0.3177	0.2177
JDMCC (Heo et al 2013)	7.4159	0.3910	0.2618	0.1894	5.1111	0.4175	0.2900	0.1962
DASC (Kim et al 2015)	7.7259	0.5485	0.3767	0.2753	7.0185	0.7290	0.5930	0.4833
CCNG (Holloway et al 2015)	8.4084	0.8453	0.7475	0.5615	8.4977	0.8490	0.7627	0.6045
ITER (Jeon et al 2016)	6.4479	0.3784	0.1572	0.1136	5.6211	0.3027	0.2397	0.1451
Learning-based								
DMC (Zhi et al 2018)	7.8583	0.7924	0.6259	0.3864	10.0653	0.8593	0.7548	0.5907
DMC + ft (Zhi et al 2018)	-	-	-	-	9.6338	0.8284	0.7063	0.5392
CMSNet (RGB pair)	4.6696	0.4576	0.2811	0.1497	4.3350 (8.0860)	0.6888 (0.7320)	0.4714 (0.5674)	0.2409 (0.3846)
CMSNet (feat. concat)	3.9881	0.3450	0.1913	0.0964	3.5442 (5.9853)	0.2306 (0.2836)	0.1547 (0.1960)	0.0820 (0.1630)
CMSNet (denoise)	4.2396	0.3476	0.1954	0.0999	3.7014 (5.5443)	0.3260 (0.3774)	0.1862 (0.2812)	0.0992 (0.1625)
CMSNet	**3.7917**	**0.2949**	**0.1656**	**0.0857**	**3.4646** (5.7505)	**0.2139** (0.2710)	**0.1448** (0.1740)	**0.0790** (0.1011)
Method	Monkaa				KITTI			
	RMSE	Bad1.0	Bad2.0	Bad4.0	RMSE	Bad1.0	Bad2.0	Bad4.0
Hand-craft matching cost								
ANCC (Heo et al 2011)	9.3956	0.5921	0.4005	0.2981	4.7791	0.3518	0.1620	0.0963
JDMCC (Heo et al 2013)	7.9282	0.5012	0.3192	0.2150	4.6405	0.2474	0.1285	0.0822
DASC (Kim et al 2015)	8.0148	0.5959	0.4097	0.2864	5.7050	0.3270	0.1965	0.1349
CCNG (Holloway et al 2015)	9.0141	0.7288	0.6166	0.4678	3.7378	0.6925	0.4352	0.1525
ITER (Jeon et al 2016)	7.1502	0.5050	0.3968	0.3004	3.8543	0.4023	0.3162	0.2248
Learning-based								
DMC (Zhi et al 2018)	9.3281	0.8740	0.7880	0.6283	7.7020	0.8425	0.7256	0.5273
DMC + ft (Zhi et al 2018)	8.6539	0.7763	0.6645	0.5248	5.9794	0.6302	0.5089	0.3178
CMSNet (RGB pair)	10.8936	0.8684	0.7709	0.6125	2.5390	0.2740	0.1548	**0.0701**
CMSNet (feat. concat)	5.4372	0.4634	0.2951	0.1828	3.0636	0.3002	0.1529	0.1002
CMSNet (denoise)	6.0270	0.5093	0.3472	0.2208	3.0736	0.3007	0.1738	0.1106
CMSNet	**4.7084**	**0.3830**	**0.2540**	**0.1578**	**2.4942**	**0.2218**	**0.1193**	0.0772

Table 2 Computational time in **disparity estimation**. Note that we utilize the official implementation codes that are provided by original authors.

Method	Inference time [sec]
Hand-craft matching cost	
ANCC (Heo et al 2011)	54.70
JDMCC (Heo et al 2013)	154.66
DASC (Kim et al 2015)	8.36
CCNG (Holloway et al 2015)	8.70
ITER (Jeon et al 2016)	120.01
Learning-based	
DMC (Zhi et al 2018)	**0.01**
CMSNet (Ours)	0.04

We also conduct an extensive ablation study to examine the effects of different components in CMSNet. We first demonstrate the benefit of the color-monochrome image pair in presence of severe noise. We train our disparity network with noisy color image pairs. To do this, we change the number of input channels in the input monochrome image from one to three, and then re-train our network. This network fails to produce reliable results because severe noise causes inaccurate matches between them. Even though the network quantitatively shows good performance on the KITTI dataset in Table 1, the estimated disparity map in Fig. 6 (9^{th} row) is blurry. The blurry disparity may come from relatively lower resolution of color images than that of monochrome images.

We compare two feature representation methods for color-monochrome pairs: correlation and concatenation. We change the cross-correlation module to a simple concatenation along with a channel dimension, and then re-train our network. In Table 1, we observe that the cross correlation provides better performance than the feature concatenation.

For another study, we slightly modify the structure of CMSNet. In this modified version, we first pass color and monochrome images through our denoising network to suppress noise, then feed the denoised images into the other subnetworks. However, its performance is worse than CMSNet because texture inconsistency between input images is increased. The basic concept of single image denoising methods (Dabov et al 2007) is to average the similar local patches, which results in edge smoothing. Since the size of the local patches is determined according to the noise levels Dabov et al (2007), images with high noise levels are severely smoothed. The CNN-based denoising methods also infer latent patches with local convolution filters. In real-world scenarios, it is infeasible for denoised images to show sharp edges and textures as same as their ground-truth images. The local convolution operations aim to smooth noisy patches to minimize RMSE over their ground-truth patches, and cause texture variations of them which make computations of matching costs difficult. As demonstrated in Hirschmüller and Scharstein (2009), Gaussian filtering, which makes noisy images blurry with the same blur kernels, is helpful for stereo matching. For better stereo matching with noisy image pairs, a novel stereo matching should be devised by considering matchability between the pairs as well as RMSE.

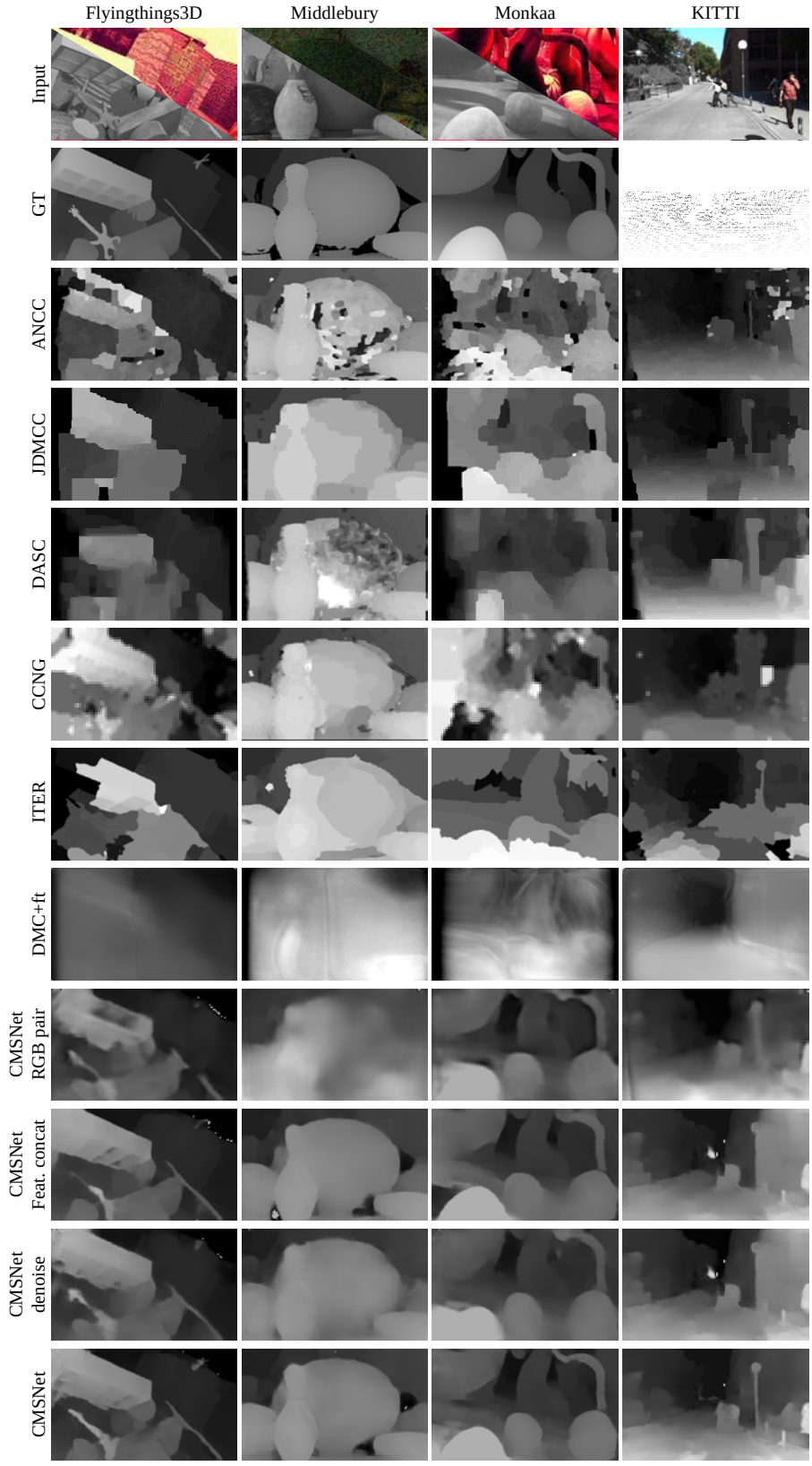


Fig. 6 Comparison of disparity map results on FlyingThings3D, Middlebury, Monkaa and KITTI. We note that the predicted disparity maps from CMSNet including ablation studies on Middlebury are results from the fine-tuned networks.

Table 3 Comparison results. **High-quality image recovery**: BM3D (Dabov et al 2007), non-local means (Buades et al 2005), DnCNN (Zhang et al 2017a), WAVG (Im et al 2019a) and ITER (Jeon et al 2016) and ours including ablation studies. The 'init. mapping' is a recovered image without the colorization process of CMSNet, and the 'w/o occlusion' is a output of the colorization network without an occlusion map.

Method	FlyingThings3D		Middlebury		Monkaa	
	PSNR (dB)	SSIM	PSNR (dB)	SSIM	PSNR (dB)	SSIM
BM3D (Dabov et al 2007)	26.0982	0.9421	30.9789	0.8288	27.3202	0.9202
Non-local (Buades et al 2005)	24.1725	0.9070	30.9424	0.8542	27.0228	0.9605
DnCNN (Zhang et al 2017a)	26.6576	0.9316	**31.0343**	0.8629	28.1681	0.9533
WAVG (Im et al 2019a)	28.2327	0.9577	30.9371	0.7757	29.5694	**0.9729**
ITER (Jeon et al 2016)	27.2488	0.9423	27.5619	0.8197	26.6232	0.9343
CMSNet (init. mapping)	27.4683	0.9377	29.1909	0.8579	29.0098	0.9512
CMSNet (w/o occlusion)	28.4535	0.9508	30.7879	0.8748	29.4817	0.9562
CMSNet	**28.5782**	**0.9598**	30.9681	**0.8749**	**29.7823**	0.9649

5.2 High-quality Color Image Recovery

We also evaluate the effectiveness of our high-quality image recovery using the FlyingThings3D, Middlebury and Monkaa datasets. For quantitative evaluation, we randomly select 50 images from each dataset and compare PSNR and SSIM (Wang et al 2004) with single image denoising and multi-image denoising methods: BM3D (Dabov et al 2007), non-local means (Buades et al 2005), DnCNN (Zhang et al 2017a), WAVG (Im et al 2019a) and ITER (Jeon et al 2016).

For the hyper-parameters of BM3D and non-local means, we utilize noise level estimations in (Liu et al 2013) and (Immerkaer 1996), respectively. DnCNN is a CNN-based single image denoising method. WAVG takes one reference image and multiple target images with their correspondences as inputs. ITER uses a color-monochrome image pair and transfers the chrominance information of the color image into the monochrome image with its corresponding disparity map. We note that ground-truth disparity maps are used for the optimal performances of WAVG and ITER in this experiment. The quantitative comparison is presented in Table 3, whose examples are shown in Fig. 7.

In the results, BM3D and non-local means show relatively higher PSNR and SSIM for the Middlebury dataset than those for Flyingthings3D and Monkaa. This is because the performances of BM3D and non-local means mainly depend on their hyper-parameters for noise levels. The noise level estimations (Immerkaer 1996; Liu et al 2013) compute the noise standard deviation from homogeneous patches of images. We observe that the noise level estimations sometimes fail to calculate optimal parameters in Flyingthings3D and Monkaa because they have highly textured images. DnCNN shows consistent performances on all datasets, but color images with high noise level lead to lower performances than that with color-monochrome image pairs. We can see that two images are insufficient for WAVG to achieve reasonable performance. ITER suffers from color bleeding error because its colorization with a chrominance consistency weight does not work in regions where parallax is significant.

We also conduct an ablation study for the color image recovery in CMSNet. We first evaluate the initial color mapping with an estimated disparity map by directly measuring PSNR and SSIM. As shown in Fig. 7 (9^{th} row), the network that does not consider the occlusion map provokes the color breeding error in out-of-plane regions. In contrast, the result of the final CMSNet in Fig. 7 (the last row) shows that the occlusion term effectively handles the artifact. In this respect, each component of CMSNet contributes to the recovery of high-quality color images. Different from conventional cascade frameworks such as BM3D, non-local means and ITER, CMSNet does not require any hyper-parameter tuning in the test phase. In addition, the recovered color images from CMSNet look perceptually convincing, thanks to the benefit of monochrome images.

5.3 Real-world results

Although we demonstrate the generality of CMSNet using the KITTI dataset in Sec. 5.1, the provided datasets were only captured under daylight conditions. For the further investigation on challenging conditions, we captured indoor and outdoor scenes with our own color-monochrome stereo camera in low-light condition and evaluated CMSNet on the real-world datasets.

We implemented our prototype system using two PointGrey Flea3 cameras, one color and one monochrome cameras, whose baseline is 5cm and the maximum disparity is about 80 pixels. The stereo system was pre-calibrated and images from the cameras were rectified using the MATLAB camera calibration toolbox[3]. A resolution of the captured images is 1200×2500 pixels, and we resize it into 567×960 pixels in the test phase. Since CMSNet consists of fully convolutional layers, images

[3] http://www.vision.caltech.edu/bouguetj/calib_doc/index.html

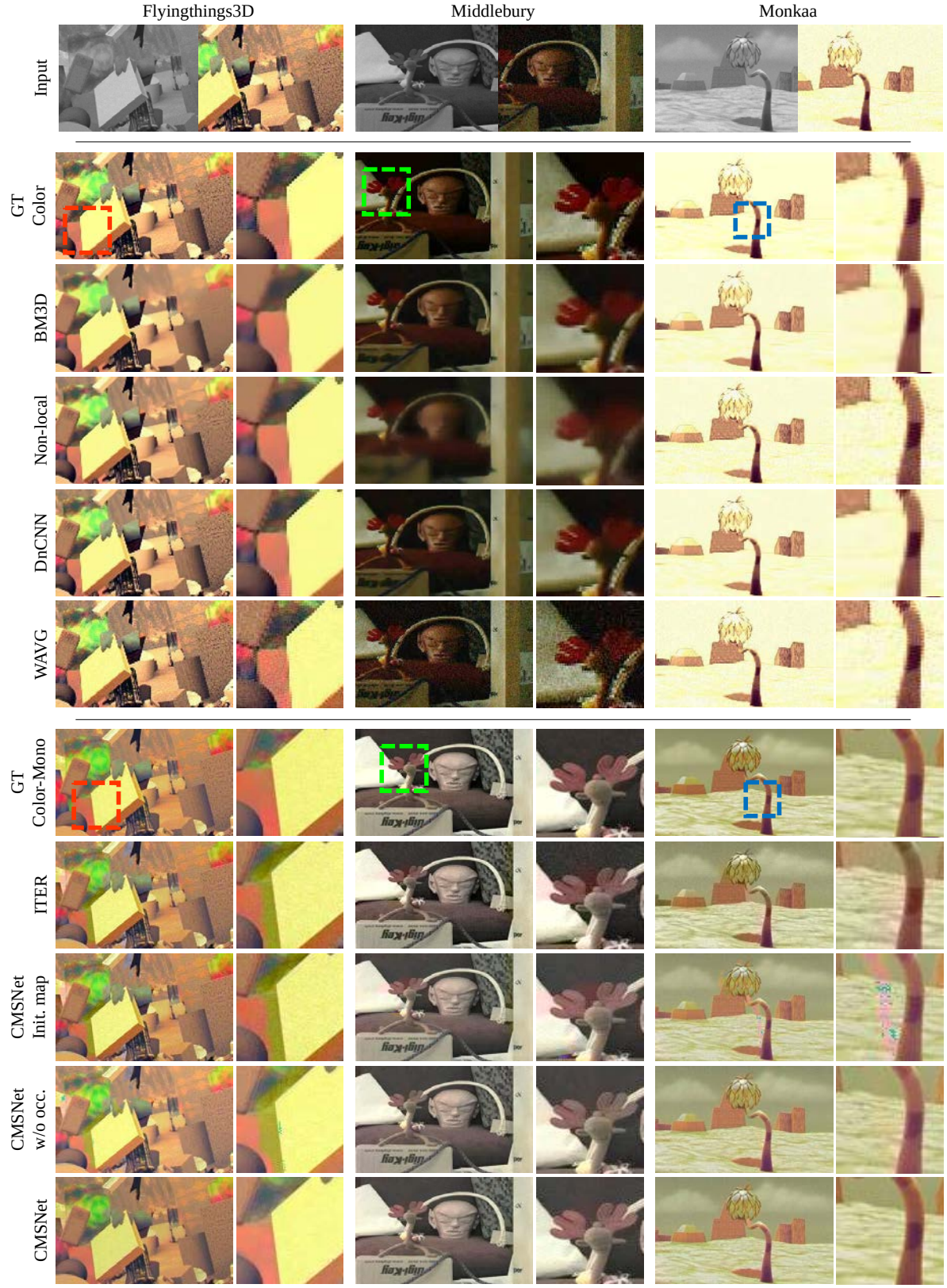


Fig. 7 Comparison of high-quality image recovery on FlyingThings3D, Middlebury and Monkaa. We categorize the results into two classes: single image denoising and merging color-monochrome pairs.

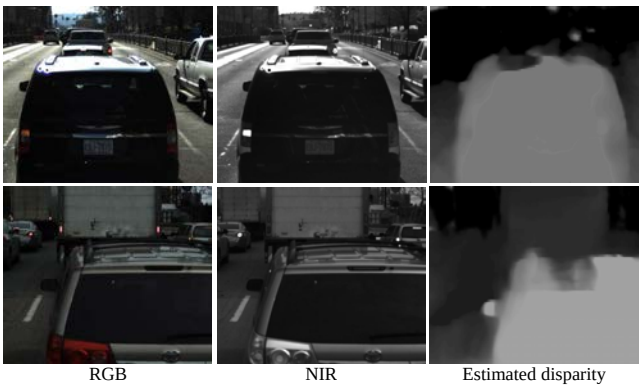
(a) Input - mono (b) Input - Color (c) Enlarged part of (a) and (d) (d) Recovered image (e) Disparity map

Fig. 8 Disparity estimation and recovered color image results on indoor/outdoor scenes captured from our prototype system.

with higher resolutions than that used in the training phase are available.

Fig. 8 shows the results of the real-world data which are captured at night in low-light conditions. CMSNet achieves accurate disparity maps and produces high-quality color images. Note that CMSNet reconstructs both depth discontinuities and fine structure such as the stone statue and the plant in the 3^{rd} row, and the golf club in the 4^{th} row of Fig. 8. Particularly, the auto-exposure of each camera does not help to recover some poorly exposed areas due to lack of light as shown in 1^{st} and 2^{nd} row of Fig. 8. On the other hand, the RGB-monochrome fusion with the estimated disparity map results in brighter images than the input color images. The use of monochrome image as a luminance channel for the recovered image enables to reduce image noise and shows higher resolution. As shown in the 4^{th} row of Fig. 8, the text are better readable after CMSNet's reasonable process.

Finally, we evaluate CMSNet on NIR-RGB stereo images[4] (Zhi et al 2018) as shown in Fig. 9 to demonstrate the versatility of the proposed method. The resolution of the dataset is 429×582 pixels and the maximum disparity is less than 20 pixels. Since the dataset does not provide ground-truth disparity maps, CMSNet trained on Flyingthings3D is just applied. As shown in Fig. 9, CMSNet shows reliable disparity estimation result without any finetuning process. Thanks to our aggressive data augmentation, our network trained on Flyingthings3D alone is enough to apply to different spectral images.



RGB NIR Estimated disparity

Fig. 9 Disparity estimation results on NIR-RGB image pairs.

6 Conclusion

We have proposed an end-to-end convolutional neural network, namely CMSNet, for high-quality disparity estimation and color image acquisition in low-light conditions. We have achieved this by utilizing a fundamental trade-off between color and monochrome cameras. We

[4] Downloaded from https://github.com/tiancheng-zhi/cs-stereo

performed extensive evaluations and validated the effectiveness of the proposed network quantitatively and qualitatively. We expect that the proposed framework can be popular as a robust stereo system for a mobile phone and a surveillance system.

In this study, we found some challenges that we should overcome, which are considered as our future work. First, CMSNet has a large number of model parameters, which is not suitable for mobile phone solutions. For a practical utility, a smaller network can be made possible by reducing redundant networks and by drilling down to architectural details. Second, the performance of CMSNet is not guaranteed for datasets with refractive media, which is still considered as a fundamental issue of stereo matching. We can introduce a material awareness term in our loss function like (Zhi et al 2018) to handle this problem.

Acknowledgment

This work is partly supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST), No.2014-3-00077, AI National Strategy Project, No.2021-0-02068, Artificial Intelligence Innovation Hub, No.2020-0-00231, Development of Low Latency VR/AR Streaming Technology based on 5G edge cloud), Vehicles AI Convergence Research & Development Program through the National IT Industry Promotion Agency of Korea (NIPA) funded by the Ministry of Science and ICT (No. S1602-20-1001), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2020 R1C1C1012635, 2020R1C1C1013210). In addition, this research was financially supported by the Ministry of Trade, Industry and Energy(MOTIE) and Korea Institute for Advancement of Technology(KIAT) through the International Cooperative RD program in part (P0019797).

References

(2016) Huawei P9. https://consumer.huawei.com/uk/phones/p9/ 1

(2017) Flea3 GigE imaging performance specification. http://www.ptgrey.com/support/downloads/10109/ 2

(2018) iphone XS. https://www.apple.com/iphone-xs/ 1

(2019) LG V50. https://www.lg.com/us/mobile-phones/v50-thinq-5g/sprint 1

(2019) Samsung Galaxy S10. https://www.samsung.com/ca/smartphones/galaxy-s10/ 1

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X (2015) TensorFlow: Large-scale machine learning on heterogeneous systems. URL https://www.tensorflow.org/, software available from tensorflow.org 7

Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Susstrunk S (2007) Multiplexing for optimal lighting. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 29(8):1339–1354 7

Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Susstrunk S (2012) Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 34(11):2274–2282 5

Buades A, Coll B, Morel JM (2005) A non-local algorithm for image denoising. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2, 11

Chakrabarti A, Freeman WT, Zickler T (2014) Rethinking color cameras. In: Proceedings of IEEE International Conference on Computational Photography (ICCP) 3

Chang JR, Chen YS (2018) Pyramid stereo matching network. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp 5410–5418 5

Dabov K, Foi A, Katkovnik V, Egiazarian K (2007) Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Transactions on Image Processing (TIP) 16(8):2080–2095 2, 9, 11

Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 3

Dong X, Li W, Wang X, Wang Y (2019) Learning a deep convolutional network for colorization in monochrome-color dual-lens system. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI) 3

Dosovitskiy A, Fischer P, Ilg E, Hausser P, Hazirbas C, Golkov V, Van Der Smagt P, Cremers D, Brox T (2015) Flownet: Learning optical flow with convolutional networks. In: Proceedings of IEEE International Conference on Computer Vision (ICCV) 5

Gallo O, Gelfandz N, Chen WC, Tico M, Pulli K (2009) Artifact-free high dynamic range imaging. In: Proceedings of IEEE International Conference on Computational Photography (ICCP) 2

Gastal ES, Oliveira MM (2011) Domain transform for edge-aware image and video processing. In: ACM Transactions on Graphics (TOG), vol 30, p 69 3, 6

Hasinoff SW, Sharlet D, Geiss R, Adams A, Barron JT, Kainz F, Chen J, Levoy M (2016) Burst photography for high dynamic range and low-light imaging on mobile cameras. ACM Transactions on Graphics (TOG) 35(6):192 2

He M, Chen D, Liao J, Sander PV, Yuan L (2018) Deep exemplar-based colorization. ACM Transactions on Graphics (TOG) 37(4):47 3

Heo YS, Lee KM, Lee SU (2011) Robust stereo matching using adaptive normalized cross-correlation. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 33(4):807–822 3, 4, 8, 9

Heo YS, Lee KM, Lee SU (2013) Joint depth map and color consistency estimation for stereo images with different illuminations and cameras. IEEE Transactions on Pattern

Analysis and Machine Intelligence (PAMI) 35(5):1094–1106 3, 8, 9
Hirschmüller H, Scharstein D (2009) Evaluation of stereo matching costs on images with radiometric differences. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 31(9):1582–1599 2, 4, 8, 9
Holloway J, Mitra K, Koppal SJ, Veeraraghavan AN (2015) Generalized assorted camera arrays: Robust cross-channel registration and applications. IEEE Transactions on Image Processing (TIP) 24(3):823–835 3, 4, 8, 9
Hu J, Gallo O, Pulli K, Sun X (2013) Hdr deghosting: How to deal with saturation? In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2
Im S, Jeon HG, Kweon IS (2019a) Robust depth estimation using auto-exposure bracketing. IEEE Transactions on Image Processing (TIP) 28(5):2451–2464 2, 11
Im S, Jeon HG, Lin S, Kweon IS (2019b) Dpsnet: end-to-end deep plane sweep stereo. In: International Conference on Learning Representations (ICLR) 5
Immerkaer J (1996) Fast noise variance estimation. Computer Vision and Image Understanding (CVIU) 64(2):300–302 11
Irony R, Cohen-Or D, Lischinski D (2005) Colorization by example. In: Eurographics Symp. on Rendering, vol 2 3
Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K (2015) Spatial transformer networks. In: Annual Conference on Neural Information Processing Systems (NeurIPS) 6
Jeon HG, Lee JY, Im S, Ha H, So Kweon I (2016) Stereo matching with color and monochrome cameras in low-light conditions. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2, 5, 6, 8, 9, 11
Kendall A, Martirosyan H, Dasgupta S, Henry P, Kennedy R, Bachrach A, Bry A (2017) End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), pp 66–75 5
Kim S, Min D, Ham B, Ryu S, Do MN, Sohn K (2015) Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 3, 8, 9
Kimmel R (1999) Demosaicing: image reconstruction from color ccd samples. IEEE Transactions on Image Processing (TIP) 8(9):1221–1228 3
Levin A, Lischinski D, Weiss Y (2004) Colorization using optimization. In: ACM Transactions on Graphics (TOG), vol 23, pp 689–694 3, 5
Li A, Yuan Z (2018) Occlusion aware stereo matching via cooperative unsupervised learning. In: Proceedings of Asian Conference on Computer Vision (ACCV) 5
Liang M, Guo X, Li H, Wang X, Song Y (2019) Unsupervised cross-spectral stereo matching by learning to synthesize. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI) 3
Liu X, Tanaka M, Okutomi M (2013) Single-image noise level estimation for blind denoising. IEEE Transactions on Image Processing (TIP) 22(12):5226–5237 11
Liu Z, Yuan L, Tang X, Uyttendaele M, Sun J (2014) Fast burst images denoising. ACM Transactions on Graphics (TOG) 33(6):232 2
Malvar HS, He Lw, Cutler R (2004) High-quality linear interpolation for demosaicing of bayer-patterned color images. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 7
Mayer N, Ilg E, Hausser P, Fischer P, Cremers D, Dosovitskiy A, Brox T (2016) A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 4, 6, 7, 8
Menze M, Geiger A (2015) Object scene flow for autonomous vehicles. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 8
Mertens T, Kautz J, Van Reeth F (2009) Exposure fusion: A simple and practical alternative to high dynamic range photography. In: Computer Graphics Forum, Wiley Online Library, vol 28, pp 161–171 2
Owen AB (2007) A robust hybrid of lasso and ridge regression. Contemporary Mathematics 443(7):59–72 5
Pinggera P, Breckon T, Bischof H (2012) On cross-spectral stereo matching using dense gradient features. In: Proceedings of British Machine Vision Conference (BMVC) 3
Quan D, Liang X, Wang S, Wei S, Li Y, Huyan N, Jiao L (2019) Afd-net: Aggregated feature difference learning for cross-spectral image patch matching. In: Proceedings of IEEE International Conference on Computer Vision (ICCV) 6
Reinhard E, Heidrich W, Debevec P, Pattanaik S, Ward G, Myszkowski K (2010) High dynamic range imaging: acquisition, display, and image-based lighting. Morgan Kaufmann 2
Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI) 6
Shen X, Gao H, Tao X, Zhou C, Jia J (2017) High-quality correspondence and segmentation estimation for dual-lens smart-phone portraits. In: Proceedings of IEEE International Conference on Computer Vision (ICCV) 1
Shin C, Jeon HG, Yoon Y, Kweon IS, Kim SJ (2018) Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 7
Tosi F, Aleotti F, Poggi M, Mattoccia S (2019) Learning monocular depth estimation infusing traditional stereo knowledge. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 5
Wang Z, Bovik AC, Sheikh HR, Simoncelli EP, et al (2004) Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing (TIP) 13(4):600–612 11
Yatziv L, Sapiro G (2006) Fast image and video colorization using chrominance blending. IEEE Transactions on Image Processing (TIP) 15(5):1120–1129 3
Zhang K, Zuo W, Chen Y, Meng D, Zhang L (2017a) Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Transactions on Image Processing (TIP) 26(7):3142–3155 2, 5, 11
Zhang K, Zuo W, Zhang L (2018) Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. IEEE Transactions on Image Processing (TIP) 27(9):4608–4622 2
Zhang R, Isola P, Efros AA (2016) Colorful image colorization. In: Proceedings of European Conference on Computer Vision (ECCV) 3
Zhang R, Zhu JY, Isola P, Geng X, Lin AS, Yu T, Efros AA (2017b) Real-time user-guided image colorization with learned deep priors. ACM Transactions on Graphics

(TOG) 9(4) 3, 6

Zhao S, Sheng Y, Dong Y, Chang EI, Xu Y, et al (2020) Maskflownet: Asymmetric feature matching with learnable occlusion mask. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 6

Zhi T, Pires BR, Hebert M, Narasimhan SG (2018) Deep material-aware cross-spectral stereo matching. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 3, 8, 9, 13, 14

7 Appendix: Details of CMSNet architecture