Low-light Color Imaging via Dual Camera Acquisition

Peiyao Guo[1][0000−0003−2887−3463] and Zhan Ma[1][0000−0003−3686−4057]

Vision Lab, Nanjing University, Nanjing, China
peiyao@smail.nju.edu.cn, mazhan@nju.edu.cn

Abstract. As existing low-light color imaging suffers from the unrealistic color representation or blurry texture with a single camera setup, we are motivated to devise a dual camera system using a high spatial resolution (HSR) monochrome camera and another low spatial resolution (LSR) color camera for synthesizing the high-quality color image under low-light illumination conditions. The key problem is how to efficiently learn and fuse cross-camera information for improved presentation in such heterogeneous setup with domain gaps (e.g., color vs. monochrome, HSR vs. LSR). We have divided the end-to-end pipeline into three consecutive modularized sub-tasks, including the reference-based exposure compensation (RefEC), reference-based colorization (RefColor) and reference-based super-resolution (RefSR), to alleviate domain gaps and capture inter-camera dynamics between hybrid inputs. In each step, we leverage the powerful deep neural network (DNN) to respectively transfer and enhance the illuminative, spectral and spatial granularity in a data-driven way. Each module is first trained separately, and then jointly fine-tuned for robust and reliable performance. Experimental results have shown that our work provides the leading performance in synthetic content from popular test datasets when compared to existing algorithms, and offers appealing color reconstruction using real captured scenes from an industrial monochrome and a smartphone RGB cameras, in low-light color imaging application.

1 Introduction

Low-light color imaging is a challenging task which plays a vital role in auto driving, security surveillance, and professional photography. Insufficient illumination which may come from the under-exposure acquisition or low-light radiation, would lead to very low signal-to-noise ratio (SNR) and corresponding severely degraded imaging quality.

Classical histogram equalization or gamma correction [1–3] was applied to directly enhance the luminance component without taking the chrominance part into account. On the other hand, as suggested in Retinex theory [4], color image could be represented by the product of its illuminance and reflectance map, where the reflectance map captures intrinsic "color" (spectral) information of the object under varying lighting conditions, and the illuminance component

describes the energy intensity of light radiation. Thus, a number of explorations had been made to decompose the illuminance and reflectance components from observed images for synthesizing better image reconstruction at a different (e.g., higher) illumination condition, where these components can be represented with either hand-crafted [5, 6] or learning-based [7–9] features. These works assumed the single camera setup in low-light condition. Though color image quality could be enhanced to some extent in this category, its reconstruction often suffered from the unrealistic color presentation, blurry texture, etc.

Recently, we have witnessed explosive advancements of multi-camera system, by which we can significantly improve the imaging capacity in various dimensions, such as gigapixel photography [10], high-speed video acquisition [11], light-field imaging [12], etc. Besides, as reported in neuronal science studies [13, 14], rods are responsible for illumination changes, especially in low-light condition, without color perception (e.g., scotopic vision), while cones are mainly for color sensation (e.g., photopic vision). Especially, the human visual system exhibits higher sensitivity to luminance variations than to chrominance under low illumination condition, since rod cells are dominantly activated (e.g., much more than cones) in such scenario [15]. All above have motivated us to apply the heterogeneous dual camera setup for low-light color imaging, where one monochrome camera is used to mimic the rod cells for capturing the monochromic image at higher spatial resolution (HSR), and the other color camera emulates the cone cells by inputting the regular color image at lower spatial resolution (LSR). Note that without requiring the color filter arrays (CFA) such as the Bayer CFA [16], monochromic imaging often provides better energy preservation of light radiance that can be leveraged to enhance corresponding color image.

Recalling the color image decomposition in Retinex theory, we have attempted to apply cross-camera synthesis by transferring the colors captured via a LSR color camera to the monochromic image acquired with a HSR monochrome camera. Because of the domain gaps in the camera pair, e.g., LSR vs. HSR, color vs. monochrome, we have proposed to divide the entire task into three consecutive sub-tasks, i.e., reference-based exposure compensation (RefEC), reference-based colorization (RefColor), and reference-based super-resolution (RefSR). Herein, RefEC downscales the HSR monochromic image to the same size of the corresponding LSR color image, and transfers the illumination level from its downscaled version to brighten the LSR color image; while RefColor module resolves the parallax between HSR and LSR cameras, and migrates the brightened LSR image colors to downscaled HSR monochromic image; In the end, re-colored and downscaled HSR image is super-resolved with the guidance of the native HSR monochromic sample in RefSR for final output. All modules, i.e., RefEC, RefColor and RefSR, are implemented using stacked convolutions to efficiently characterize and learn the illumination-, spectrum (color)- and resolution-dependent dynamics between proposed hybrid inputs. We first train each modularized component individually, and then fine-tune the end-to-end pipeline for robust and reliable low-light imaging.

Experimental results have demonstrated that our method shows leading performance in each task using popular datasets when compared with relevant algorithms. Simulations have also revealed appealing image reconstruction using camera-captured real scenes under low-light illumination conditions. Overall, main contributions of this work are summarized below:

- We are motivated to devise a dual camera system for low-light color imaging, where a HSR monochromic image and a LSR color image from respective cameras are synthesized for final enhanced color image under low-light illumination condition; Such heterogeneous camera setup is inspired by the non-uniform light responses of retinal rods and cones for luminance and chrominance.
- We have divided the entire system into three consecutive sub-tasks, i.e., RefEC, RefColor and RefSR, by implicitly enforcing the cross-camera reference to alleviate domain gaps and capture cross-camera dynamics (e.g., illumination, spectrum, spatial resolution) for synthesizing the final high-quality output.
- Our method shows competitive performance on both public datasets and the real captured scenes, promising the generalization in practical applications.

2 Related work

This work is closely related to the multi-camera imaging, low-light image enhancement, colorization and super-resolution. A brief review is given below.

Multi-camera Imaging. As aforementioned, multi-camera system could significantly improve the imaging capacity by computationally synthesizing input sources for gigapixel photography, high-speed videography, light-fields, hyperspectral imaging, etc. Wang *et al.* [17] have proposed to register a pair of RGB and NIR-G-NUV image for addressing the motion blur and temporal delay of dark flash photography. Trinidad *et al.* [18] have applied an end-to-end feature fusion from multiple misaligned images for high-quality image generation in color transferring, high dynamic range imaging, texture restoration, etc. Dong *et al.* [19] have imposed a monochrome and color dual-lens setup to shoot high-quality color images. Recently, multi-camera system becomes a commodity and is widely adopted in mobile platforms for super-resolution [20, 21], denoising [22] and quality enhancement [23].

Low-light Image Enhancement. In this category, classical approaches include the histogram equalization (HE) and gamma correction [1, 3]. However, they fail to retain the local details and suppress noise. Leveraging the characteristics of low-light image, dehazing model [24, 25] assumes that the low-light image resembles the haze image after inversion, and Retinex model [26–29, 9, 8] decomposes the reflectance map from the observed image for synthesizing it at a higher illumination condition. For example, Wei *et al.* [29] reconstruct the high-quality image with the decomposed reflectance map and enhanced illuminance map using an end-to-end learning approach. And Zhang *et al.* [8] assume

the histogram consistency after equalization for low-light image enhancement. Besides, Guo *et al.* [30] impose a non-reference enhancement for dynamic range adjustment via high-order polynomial function-based pixel-wise processing.

Colorization. Automatic colorization, scribble-based colorization and exemplar-based colorization are the main types of colorization methods. Abundant automatic proposals [31–34] benefit from the supervised colorization training on large datasets. Scribble-based methods [35, 36] focus on propagating local user hints to the entire monochrome image. These methods are prone to produce visual artifacts such as chromatic aberration since the color priors heavily depend on the training dataset or user preference. Exemplar-based colorization provides a similar reference for the input monochrome image from pixel [37] or semantic level [38, 39], to generate more plausible colors without manual effort. Dong *et al.* [37] have utilized weighted average of colors of all potential pixels in the reference image to approximate correct color. He *et al.* [38] have proposed to use VGG-19 feature of gray-scale image to measure the semantic similarity between the reference and target image for color propagation while Zhang *et al.* [39] have added contextual loss to semantically constrain the region re-sampling in intra-colorization.

Super-resolution. Super-resolution techniques have been widely utilized in applications. Learning-based methods have been dominantly leveraged for single image super-resolution (SISR) because of its superior performance, such as [40–42]. Recent explorations have then introduced another reference image (e.g., often from an alternative camera or from a semantically similar scene) as the prior to further improve SR performance. This is so-called Reference-based super-resolution (refSR). CrossNet [12] and AWnet [11] use the refSR to generate high-quality images in light-fields imaging and high-speed videography. Additionally, Yang *et al.* [43] suggest the SR improvement with finer texture reconstruction by learning the semantic priors from the high-definition reference.

3 Method

3.1 Framework

Our system is generally shown in Fig. 1, where we input a HSR monochrome image and another LSR color image for final HSR color image reconstruction. To efficiently characterize the cross-domain variations (e.g., illumination, spectra and resolution) for better synthesis, we have divided the entire pipeline into modularized RefEC, RefColor and RefSR consecutively. Each module is implemented with DNNs for massively exploiting the power of data-driven learning methodology.

Let l be the light energy or illumination intensity, s as the spatial resolution, v as the viewpoint, and Y, U, V as respective color/spectral components[1]. The

[1] Here, Y and UV represent the luminance and chrominance components in YUV color space that is widely adopted in image/video applications.

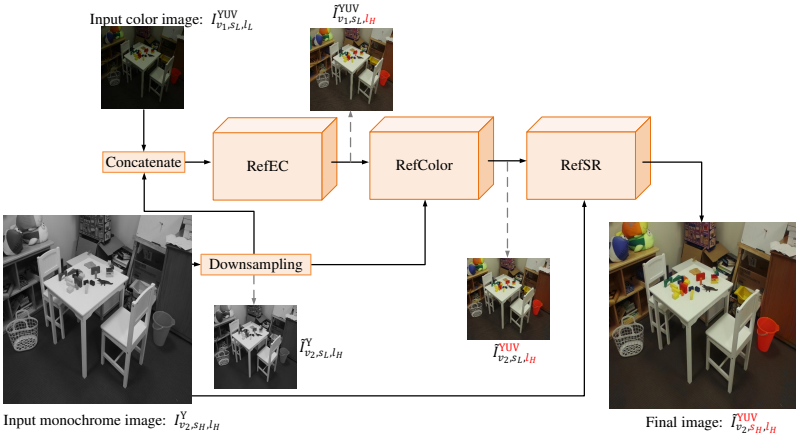

Fig. 1. Framework. A cascaded workflow inputs a HSR monochrome image and a LSR color image for final HSR color image reconstruction using modularized RefEC, RefColor and RefSR to characterize, learn and fuse cross-camera information (e.g., illumination, spectra, and resolution). The intermediate outputs are also provided using dash lines for step-wise illustration.

input LSR color image is formulated as $I^{\mathrm{YUV}}_{v_1,s_L,l_L}$, while the HSR monochromic image from another camera is $I^{\mathrm{Y}}_{v_2,s_H,l_H}$. Normally, monochrome camera offers better light radiance preservation with higher illumination intensity, and higher spatial resolution without Bayer sampling than corresponding color camera. Thus, we simply use subscripts H and L (a.k.a., high and low) to indicate the difference.

In low-light condition, it is an ill-posed problem to reconstruct high-quality color information from a single image due to insufficient exposure prior. Recent learning-based computational imaging [44] motivates us to fuse cross-camera characteristics for high-quality color image reconstruction under low-light illumination condition. Considering that the human visual system is less sensitive to the chrominance than the luminance component, we suggest to transfer the colors from a LSR color image to anther HSR monochrome image in a dual camera system. First, RefEC learns the light radiation level from downscaled monochromic image $I^{\mathrm{Y}}_{v_2,\tilde{s}_L,l_H}$ that is the same size scale of $I^{\mathrm{YUV}}_{v_1,s_L,l_L}$, and compensates $I^{\mathrm{YUV}}_{v_1,s_L,l_L}$ to $\tilde{I}^{\mathrm{YUV}}_{v_1,s_L,l_H}$ with brighter color; And in RefColor module, the color information is transferred from $\tilde{I}^{\mathrm{YUV}}_{v_1,s_L,l_H}$ at v_1 to $I^{\mathrm{Y}}_{v_2,\tilde{s}_L,l_H}$ at v_2 assuming that the similar luminance component shall have close chrominance intensity [37], resulting in $\tilde{I}^{\mathrm{YUV}}_{v_2,s_L,l_H}$. Nevertheless, it often incurs the missing regions due to the parallax induced occlusion; Thus an additional post refinement block is included in RefColor to improve warped chrominance components. Finally, $\tilde{I}^{\mathrm{YUV}}_{v_2,s_L,l_H}$ is interpolated to native higher resolution as the raw HSR monochromic input, leading to the final output $\tilde{I}^{\mathrm{YUV}}_{v_2,s_H,l_H}$. Our pipeline stepwisely learns and aggregates the dynamics of input LSR color and HSR monochromic images for robust and reliable reconstruction. More details are introduced in the following sections.

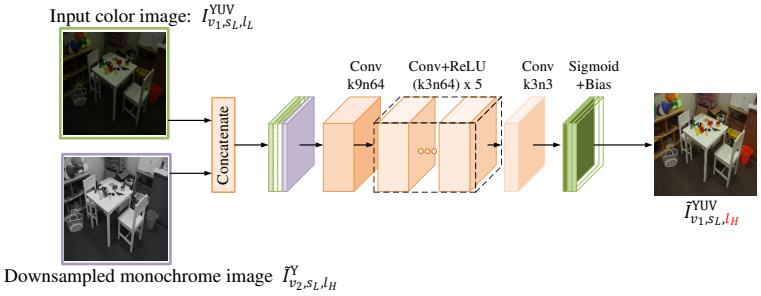

Fig. 2. RefEC. k9n64 indicates a kernel size of 9×9 and a feature map channel number of 64. Similar conventions are applied to k3n64, k3n3.

3.2 RefEC Net

Retinex Theory [4] is widely used in the image enhancement task to reconstruct real scenes. It assumes the captured color image I could be represented with the element-wise product of illuminance and reflectance map, denoted as below,

$$I = L \cdot R \tag{1}$$

Here, the illuminance map L describes the overall light condition which depends on light source, sensor quantum efficiency and integration time. (To distinguish the illuminance map here with the luma component in YUV color space, we describe the Y channel feature with the brightness or intensity.) The reflectance map R depicts the object's intrinsic color information which keeps constant in various light condition. Commonly, R consists of spectral components like RGB channels for color representation. When I denotes a monochrome image, the decomposition could be in the following form.

$$I = L \cdot \sum_{i \in C} R_i \tag{2}$$

According to Eq. (1), (2), to enhance the image which obtained with pool photon conversion, the L component of $\tilde{I}^{YUV}_{v_1,s_L,l_H}$ could draw on that of $I^{Y}_{v_2,\tilde{s}_L,l_H}$ which shares similar perspective. Inspired by the decomposition module in [7], we handle this problem with stacked CNN modules. Instead of separating exact illuminance and reflectance map, we directly obtain the predicted images following the reference's light condition.

As shown in Fig. 2, the input of RefEC net is a dual color-monochrome image pair at the lower resolution, namely, $I^{YUV}_{v_1,s_L,l_L}$ and the downscaled $I^{Y}_{v_2,\tilde{s}_L,l_H}$ on the basis of $I^{Y}_{v_2,s_H,l_H}$. RefEC module is made up of one 9×9 and six 3×3 convolutional layers. Most convolutional layers are followed by a ReLU layer with the exception of the first and last ones. The first convolutional layer extracts features with a large receptive field. Successive layers exploit the non-linearity to establish high dimension characteristics. And a sigmoid function and constant bias follows the last convolutional layer to limit $\tilde{I}^{Y}_{v_1,s_L,l_H} \in [0,1]$ and $\tilde{I}^{UV}_{v_1,s_L,l_H} \in [-0.5,0.5]$.

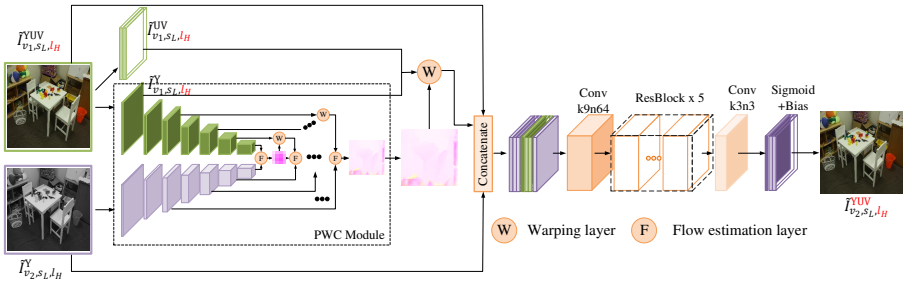

Fig. 3. RefColor. Resblock consists of a k3n64 convolutional layer followed by a batch norm and ReLU layer. PWCNet is used for flow-based color correlation measurement.

3.3 RefColor Net

The goal of the RefColor module is to colorize the coarse monochromic image $I^{\mathrm{Y}}_{v_2,\tilde{s}_L,l_H}$ based on $\tilde{I}^{\mathrm{YUV}}_{v_1,s_L,l_H}$. As the input dual pairs share the similar perspective, conventional methods commonly adopt hand-crafted features like Gabor features, SURF descriptors or DCT tansformation [45, 46] to perform a long match with spatial priors, while optical flow could quickly record pixel-level motion in frame prediction and stereo matching tasks [47–50]. Hence, we measure the color correlation based on the optical flow between $I^{\mathrm{Y}}_{v_2,\tilde{s}_L,l_H}$ and $\tilde{I}^{\mathrm{Y}}_{v_1,s_L,l_H}$. Here, we empirically take PWCNet [51] as optical flow estimation backbone due to its compact and efficient feature representations. Even though warping $\tilde{I}^{\mathrm{YUV}}_{v_1,s_L,l_H}$ to the viewpoint v_2 could restore most color information, there are still no appropriate matches for some occluded areas. To fill these holes in the rough result, we add residual-block modules to fuse all reference images and predict each pixel's color information in an end-end way.

As illustrated in Fig. 3, the Y channel of $\tilde{I}^{\mathrm{YUV}}_{v_1,s_L,l_H}$ and $I^{\mathrm{Y}}_{v_2,\tilde{s}_L,l_H}$ are fed into PWCNet module for reference pixel prediction. To make the best of features at different granularity, PWCNet progressively links different pyramid-level features with the warping layer to estimate large displacement flow. The coarse-to-fine concept could weaken the effect of large parallax but there still exists some missing areas. After warping $\tilde{I}^{\mathrm{YUV}}_{v_1,s_L,l_H}$ to v_2, the rough color map$\tilde{I}^{\mathrm{YUV}}_{v_2,s_L,l_H}$ is fused with the monochrome image $I^{\mathrm{Y}}_{v_2,\tilde{s}_L,l_H}$ and the referenced color image $\tilde{I}^{\mathrm{YUV}}_{v_1,s_L,l_H}$ through residual blocks to directly predict each pixel's color information $\tilde{I}^{\mathrm{YUV}}_{v_2,s_L,l_H}$. Note that sigmoid activation and constant bias are also applied here to make the results follow the YUV space limitation.

3.4 RefSR Net

On the strength of convolution kernel's efficient feature representation, the coarse color information $\tilde{I}^{\mathrm{YUV}}_{v_2,s_L,l_H}$ is obtained via tranferring $\tilde{I}^{\mathrm{YUV}}_{v_1,s_L,l_H}$'s chrominance to $I^{\mathrm{Y}}_{v_2,\tilde{s}_L,l_H}$. To reconstruct the high-definition color images, we need to interpolate $\tilde{I}^{\mathrm{YUV}}_{v_2,s_L,l_H}$ to the higher resolution. The HSR monochromic input $I^{\mathrm{Y}}_{v_2,s_H,l_H}$ reserves

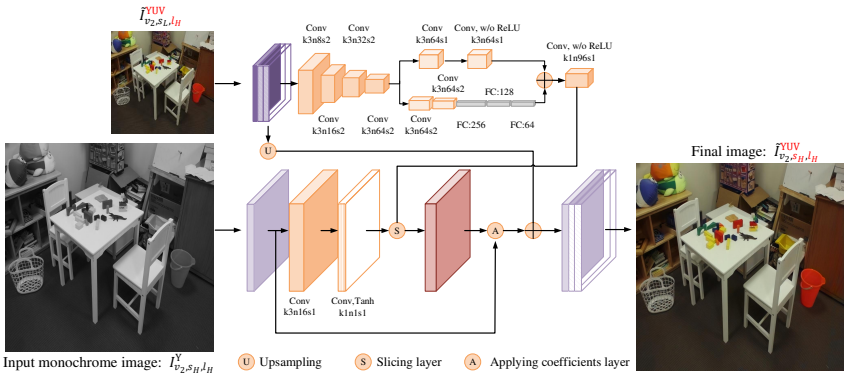

Fig. 4. RefSR. k3n8s2 indicates the convolutional layer with a kernel size of 3×3, a feature map channel number of 64 and a stride of 2. FC,256 denotes the full connection layer with the output channel size of 256. Most convolutional layers are followed by ReLU and the exceptions are clearly annotated. Slicing layer is a tri-linear interpolation operator considering space and intensity effect proposed in [53]. And in the applying coefficients layer, the monochrome image is element-wisely multiplied with the upsampled color coefficients.

the complete structure information which could suppress oversoomth effect in the chrominance interpolation although without color decomposition. Since the trainable slicing layer in [52] shows impressive performance in edge preservation, we take the similar network as HDRNet [52] to reconstruct finer chrominance information.

As shown in Fig. 4, RefSR net extracts color and structure representations from the LSR color image and HSR monochromic image respectively. Instead of directly interpolating $\tilde{I}^{\mathrm{YUV}}_{v_2,s_L,l_H}$ with the guidance of $I^{\mathrm{Y}}_{v_2,s_H,l_H}$ pixel by pixel, this module decomposes the scene's chrominance at the low resolution and then interpolates the color coefficients to the high resolution in a bilateral-grid upsampling way with the $I^{\mathrm{Y}}_{v_2,s_H,l_H}$'s guidance. In the low-resolution branch, the input $\tilde{I}^{\mathrm{YUV}}_{v_2,s_L,l_H}$ is firstly converted into the representation of multiple channels at lower resolution after cascaded convolutional layers' processing, which decides the channel granularity for successive bilateral grid. The following local and global stream take the surrounding pixels and the overall consistency into consideration and combine with each other for coarse color coefficients to alleviate large variations in the flat region. To reconstruct finer color coefficients at the high resolution, the HSR monochrome image is projected into the channel space to guide the bilateral-grid upsampling of coarse color coefficients. For better chrominance interpolation, we impose two convolutional layers in the UV space following the fusion of color coefficients and the monochrome image. Afterwards, we add the color residual to the upsampled coarse one for high-quality color images.

4 Experiments

We divide the low light image enhancement task into three independent subtasks. We randomly crop image patches from various datasets for different subtask. And in our task, the dual monochromic-color image pair is simulated via images at different specification as shown in Table 1. The input of our models are converted into the YUV color space [54] and the Y channel feature is chosen as the monochromic input. Besides, to imitate the capture noise in real scenes, various amount of noises to are added to both training and test image pairs as recommended in [55, 56]. Due to the page limitation, we provide more training details in the supplementary.

Table 1. Details on the training data setup

Task	Datasets	Viewpoint	Scale	Light energy numbers	
RefEC	Middleburry2006	(v_1, v_2)	(s_L, s_L)	(l_L, l_H)	45K
RefColor	FlyingThings3D	(v_1, v_2)	(s_L, s_L)	(l_H, l_H)	26K
RefSR	DIV2K	(v_2, v_2)	(s_L, s_H)	(l_H, l_H)	15K
Overall	Middleburry2006	(v_1, v_2)	(s_L, s_H)	(l_L, l_H)	1596
Description	The parameters in $(\cdot, \cdot)$ denote the specification of input color and monochromic image respectively. $(v_1, v2)$ represents different viewpoints. s_L denotes the spatial resolution at 256×256 while $s_H = 1024 \times 1024$. l_L, l_H means different light energy obtained by different sensors. Here we leverage various exposure time to imitate light efficiency diversity between monochrome and color sensors [56].				

Moreover, we also evaluate our model in other datasets and compared with different proposals. Our work is also validated in the monochrome-color smartphone camera module. More details are shown in the following section 4.2.

4.1 Implementation

Our framework is implemented in PyTorch on NVIDIA GTX1080 GPU. All models are optimized with Adam [57]. L1, MSE and Cosine similarity loss are used to supervise the information reconstruction in the YUV space. We pretrain three subtasks with different datasets and than finetune the whole pipeline with Middleburry2006 dataset for more robustness. L1 and Cosine similarity loss for refEC and refColor task at the lower resoltuion are beneficial to sharp details preservation. And MSE loss dominates in the refSR task for faster convergence. The initial learning rate (lr) of refEC and refSR module is set to $1e^{-4}$. As we exploit the officially pretrained PWCNet model for refColor module initialization, the lr for the gray correlation and refinement module decay from $1e^{-5}$ and $5e^{-5}$,

respectively. Afterwards, we finetune the whole pipeline with batchsize = 24, lr = $5e^{-}5$ (except $1e^{-5}$ for the gray correlation module) based on each module's pretrained model. The weight ratio for L1 and Consine similarity is 1 and 0.1 for the finetune procedure.

4.2 Comparison

In this part, we compare our individually-trained modules with other public works in each subtask and then give the performance of the whole pipeline on the simulated and real-captured scenes. Further user study and ablation study are introduced in the supplementary.

RefEC As the previous work introduced, low light image enhancement and histogram equalization (HE) both can adjust the image's intensity to the normal level. We pick up HE, Retinex Net [7], SSIE Net [8], Zero-DCE [30] methods as the reference result since we leverage different dimension information to handle this problem. Middleburry2014 [58] and S7 ISP Dataset [59] are chosen as test datasets which contain the indoor and outdoor underexposure cases. As shown in Table 2, our work shows great performance above other methods. Note that other methods only use a single frame to enhance image light condition and it is a tough problem to estimate good illumination level. Fig. 5 presents qualitative results of these methods. HE only adjusts the brightness to uniformly distribute among the whole range and couldn't compensate for chroma information. Retinex net reconstructs sharp textures but loses realistic gloss on the captured surface. Based on Retinex Net, SSIE Net still suffers from the absence of realistic surface gloss even though it shows high brightness. Zero-DCE takes color constancy as well as local exposure into consideration for good reconstruction, but suffers from some ringing artifacts around the edge due to the smooth illuminance constraint. Our work exploits the monochromic image's illumination as the guidance to adjust the underexposure image to the normal light level, which preserves sharpness and compensate chrominance meanwhile. It is manifested via the increase on PSNR_{UV} in contrast to the input. But the reconstructed highlight areas are relatively dim due to the global intensity consistency.

RefColor In this task, we collect different proposals from automatic colorization [34] (IAIC), exemplar colorization [38] (DEC) and conventional image patch match [60] (PM) to illustrate our work. The test datasets consist of indoor and outdoor scenes from Middleburry2014 and Cityscapes [61]. Since the input images are captured from different viewpoint, it imposes huge pressure to find the appropriate reference patch with global traversal, especially in the strong contrast region like the third scene in Fig 6. The performance of automatic ICIA depends on the instance segmentation and it may fails to reconstruct colors when the instance is not accurately detected. DEC leverages semantic features to search the potential match in the reference. But this match may lose the

Table 2. Quantitative average results of exposure compensation models on Middleburry2014 and S7 ISP datasets.

Model	$PSNR_{YUV}$	$PSNR_{UV}$	MS-SSIM	$PSNR_{YUV}$	$PSNR_{UV}$	MS-SSIM
HE	15.1593	29.31	0.6991	20.7308	28.5491	0.858
Retinex Net [7]	17.74	24.51	0.719	20.58	25.689	0.76
SSIE Net [8]	16.5245	24.6488	0.8422	18.2423	22.1694	0.8473
Zero-DCE [30]	26.86	34.7532	0.9327	24.3149	32.7543	0.9271
Ours	33.38	38.45	0.981	26.81	32.97	0.95
input-label	17.301	29.42	0.655	14.99	28.64	0.700
Dataset	Middleburry2014			S7 ISP		

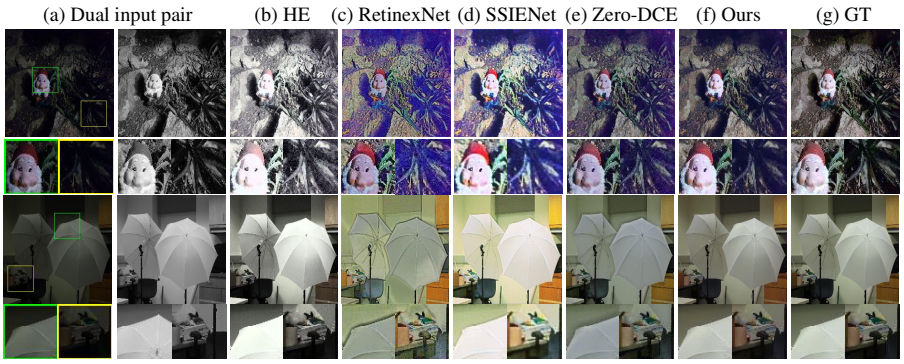

Fig. 5. Qualitative results of different image enhancement algorithms:(a) are the input color image with low luminance and the reference monochrome image. (g) is the ground truth with normal luminance.

awareness of the object structure and result in color distortion around the object boundary. In our work, Y channel features which preserve complete structure information are exploited to calculate the correlation between the target and reference image and shows better performance in colorizing details. However, similarly, when the gray channel correlation is not well measured , color bleeding will affect the colorization's quality.

RefSR There are many proposals in the field of the super resolution. The usage of deep learning benefits texture reconstruction in both single frame or multiple frame super resolution. Most of them show the impressive performance and we list some typical models like RCAN [42], TTSR [43] for the comparison. Here, we use the captured low-high resolution image datasets City100 [62] to validate the models' performance. We adopt the pretrained model in their origin work to process the low resolution color image. Note that we use TTSR model which is trained with only reconstruction loss for higher quantitative performance and

Table 3. Quantitative average results of colorization models on Middleburry2014 and Cityscapes datasets. We randomly choose 100 cases from Cityscapes dataset for test due to PM's low computation efficiency.

Model	$PSNR_{YUV}$	$PSNR_{UV}$	MS-SSIM	$PSNR_{YUV}$	$PSNR_{UV}$	MS-SSIM
PM [60]	29.6468	29.0148	0.9272	33.7849	35.6041	0.9796
IAIC [34]	28.8133	27.0774	0.9455	-	-	-
DEC [38]	37.0040	36.0793	0.9825	39.2985	38.7194	0.9927
Ours	41.0477	39.48	0.9924	42.1154	40.791	0.9968
Dataset	Middleburry2014			Cityscapes		

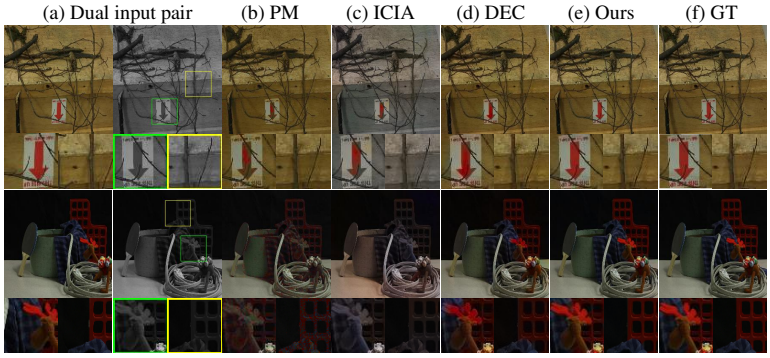

Fig. 6. Colorization results of different proposals.

slice the test image into small patches as the input of TTSR due to GPU memory shortage. Table 4 demonstrates the results of various super resolution methods. In the sharp edges or stripe structure regions, RCAN's result is more blurry than others since it only takes the single low resolution frame as the input. It is also demonstrated that our proposal preserves more clear textures within color interpolation when compared to the information alignment with soft and hard attention mechanism in TTSR.

Table 4. Quantitative results of super resolution models on City100 datasets.

Model	$PSNR_{YUV}$	$PSNR_{UV}$	MS-SSIM	$PSNR_{YUV}$	$PSNR_{UV}$	MS-SSIM
RCAN [42]	29.1734	38.226	0.8924	31.4657	36.3924	0.9099
TTSR-rec [43]	29.365	38.2135	0.8965	32.7906	35.4508	0.9234
Ours	40.52	39.066	0.9870	38.6833	37.0752	0.9755
Dataset	City100-iphone			City100-Nikon		

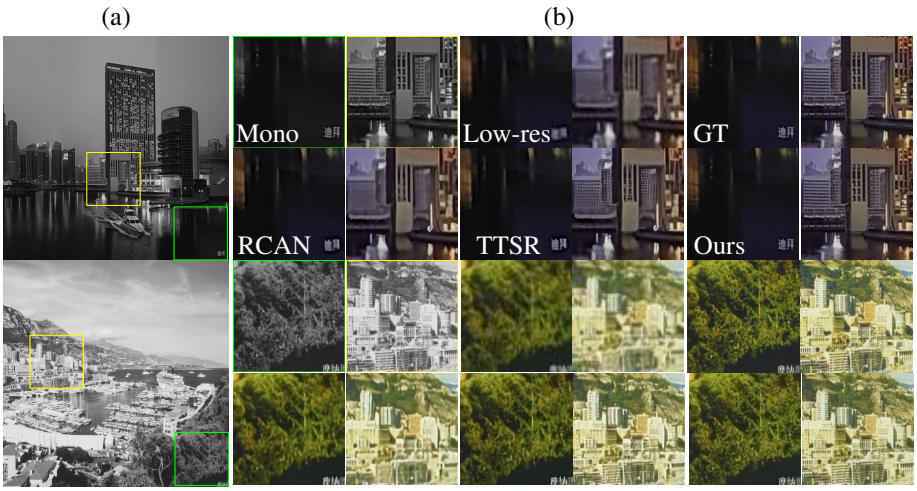

Fig. 7. Qualitative results of super resolution proposals: (a) represents the input monochrome image at the high resolution. And another color image is at the 14×14 scale of it. (b) illustrates the details of corresponding green and yellow patches in respectively, monochrome image, color image ($\times 4$ for display), ground truth, RCAN, TTSR and our method.

Overall pipeline For robustness validation, we execute the test on the extra simulated datasets and some monochrome-color pairs captured via the dual cameras. We resize the image pairs of Middleburry2014 into 256×256 and the corresponding 1024×1024 scale. And in this simulated datasets, the average PSNR_{YUV}, PSNR_{UV} and MS-SSIM of overall model reach to **38.604dB, 36.906dB** and **0.9804** respectively, while the corresponding performances of the cascaded separative model are 38.444dB, 36.816dB and 0.9802. Besides, we also use the monochrome-color industrial pair cameras and HuaweiP20 to capture the real scenes. In the industrial cameras case, we turn the aperture to the minimum ($F = 16$) and capture the objects in the normal light condition with the shutter speed of 30ms. We pre-crop the captured images to generate the $4\times$-$1\times$ monochrome-color image pair. While in HuaweiP20 case, the smartphone automatic exposures with the only color or monochrome sensor. As shown in Fig. 8, our proposal successfully transfers the chrominance of the LSR image to the HSR monochromic image with the noise suppressed.

As we finally refine the overall enhancement in an end-to-end way with the only supervision of re-colored HSR monochromic image, performance on each subtask is affected by global optimization which is shown in Fig.9, when compared with cascaded individually-trained models. And error accumulation becomes severe due to the cascaded mechanism, especially when color transfer fails from the color viewpoint to the monochrome viewpoint at the low resolution.

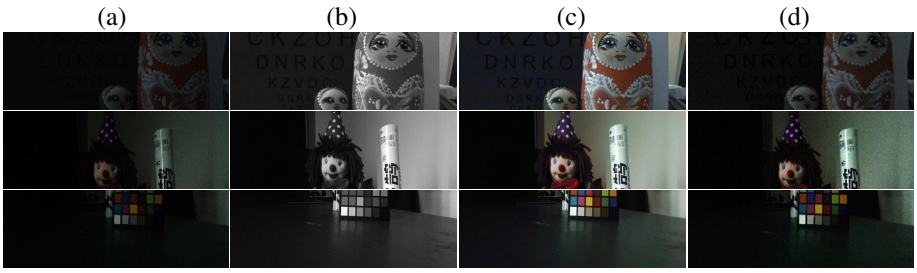

Fig. 8. The performance on the real scenes: Here, we only display partial regions due to space shortage. (a) denotes the LSR color image (×4 scale). (b) and (c) are the HSR monochromic image as well as its color reconstruction. (d) is the value-amplified LSR image for showing noise and blur (×4 scale).

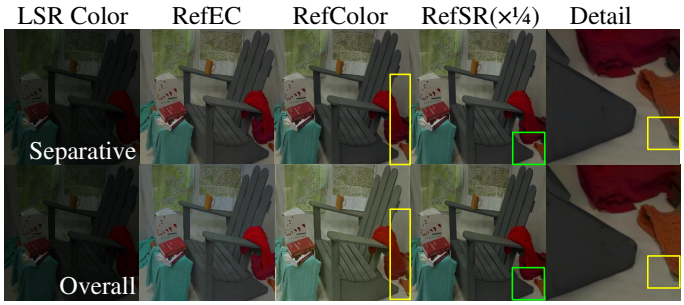

Fig. 9. The performance of the separative and overall model: In the overall model, the result of each subtask has obvious color aberration without hidden supervision but the final generation does well in details. We enlarge the green patch in 4th column to show the difference.

5 Conclusion

We present a cost-effective dual-camera system for low-light color imaging with a HSR monochromic camera and a LSR color one. Such end-to-end cross-camera synthesis is decomposed into consecutive reference-based exposure compensation, reference-based colorization and reference-based super resolution, by which we can effectively capture, learn and fuse hybrid inputs for high-quality color image with improved granularity of illumination, spectra and resolution. Extensive experiments using both synthetic images from public datasets and real captured scenes evidence that our work offers the encouraging low-light imaging efficiency with such dual camera setup.

6 Acknowledgement

We are grateful for the constructive comments from anonymous reviewers. The corresponding author is Dr. Zhan Ma (mazhan@nju.edu.cn).

References

1. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B.M., Zimmerman, J.J.: Adaptive histogram equalization and its variations. Graphical Models graphical Models and Image Processing computer Vision, Graphics, and Image Processing **39** (1987) 355–368
2. Coltuc, D., Bolon, P., Chassery, J.M.: Exact histogram specification. IEEE Transactions on Image Processing **15** (2006) 1143–1152
3. Ibrahim, H., Kong, N.S.P.: Brightness preserving dynamic histogram equalization for image contrast enhancement. IEEE Transactions on Consumer Electronics **53** (2007)
4. Land, E.H.: The retinex theory of color vision. Scientific American **237 6** (1977) 108–28
5. Wang, S., Zheng, J., Hu, H.M., Li, B.: Naturalness preserved enhancement algorithm for non-uniform illumination images. IEEE Transactions on Image Processing **22** (2013) 3538–3548
6. Fu, X., Zeng, D., Huang, Y., Zhang, X.P.S., Ding, X.: A weighted variational model for simultaneous reflectance and illumination estimation. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2782–2790
7. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. ArXiv **abs/1808.04560** (2018)
8. Zhang, Y., Di, X.G., Zhang, B., Wang, C.: Self-supervised image enhancement network: Training with low light images only. ArXiv **abs/2002.11300** (2020)
9. Wang, J., Tan, W., Niu, X., Yan, B.: Rdgan: Retinex decomposition based adversarial learning for low-light enhancement. 2019 IEEE International Conference on Multimedia and Expo (ICME) (2019) 1186–1191
10. Brady, D.J., Pang, W., Li, H., Ma, Z., Tao, Y., Cao, X.: Parallel cameras. Optica **5** (2018) 127–137
11. Cheng, M., Ma, Z., Asif, S., Xu, Y., Liu, H., Bao, W., Sun, J.: A dual camera system for high spatiotemporal resolution video acquisition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
12. Zheng, H., Ji, M., Wang, H., Liu, Y., Fang, L.: Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In: ECCV. (2018)
13. Wikipedia contributors: Retina — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Retina&oldid=964207154` (2020) [Online; accessed 30-June-2020].
14. Morie, J., McCallum, K.: Handbook of Research on the Global Impacts and Roles of Immersive Media. Advances in Media, Entertainment, and the Arts. IGI Global (2019)
15. Robinson, S., Schmidt, J.: Fluorescent penetrant sensitivity and removability: What the eye can see, a fluorometer can measure. Materials evaluation **42** (1984) 1029–1034
16. Bayer, B.E.: Color image array. US Patent **3971056** (1976)
17. Wang, J.J., Xue, T., Barron, J.T., Chen, J.: Stereoscopic dark flash for low-light photography. 2019 IEEE International Conference on Computational Photography (ICCP) (2019) 1–10
18. Trinidad, M.C., Martin-Brualla, R., Kainz, F., Kontkanen, J.: Multi-view image fusion. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 4100–4109

19. Dong, X., Li, W.: Shoot high-quality color images using dual-lens system with monochrome and color cameras. Neurocomputing **352** (2019) 22–32
20. Chu, X., Zhang, B., Ma, H., Xu, R., Li, J., Li, Q.: Fast, accurate and lightweight super-resolution with neural architecture search. arXiv preprint arXiv:1901.07261 (2019)
21. Wronski, B., Garcia-Dorado, I., Ernst, M., Kelly, D., Krainin, M., Liang, C.K., Levoy, M., Milanfar, P.: Handheld multi-frame super-resolution. ACM Transactions on Graphics (TOG) **38** (2019) 1–18
22. Godard, C., Matzen, K., Uyttendaele, M.: Deep burst denoising. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 538–554
23. Ignatov, A., Van Gool, L., Timofte, R.: Replacing mobile camera isp with a single deep learning model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. (2020) 536–537
24. Dong, X., Wang, G., Pang, Y., Li, W., Wen, J., Meng, W., Lu, Y.: Fast efficient algorithm for enhancement of low lighting video. In: ICME. (2011)
25. Li, L., Wang, R., Wang, W., Gao, W.: A low-light image enhancement method for both denoising and contrast enlarging. 2015 IEEE International Conference on Image Processing (ICIP) (2015) 3730–3734
26. Kimmel, R., Elad, M., Shaked, D., Keshet, R., Sobel, I.: A variational framework for retinex. International Journal of Computer Vision **52** (2004) 7–23
27. Fu, X., Zeng, D., Huang, Y., Ding, X., Zhang, X.P.S.: A variational framework for single low light image enhancement using bright channel prior. 2013 IEEE Global Conference on Signal and Information Processing (2013) 1085–1088
28. Guo, X., Li, Y., Ling, H.: Lime: Low-light image enhancement via illumination map estimation. IEEE Transactions on Image Processing **26** (2017) 982–993
29. Chen Wei, Wenjing Wang, W.Y., Liu, J.: Deep retinex decomposition for low-light enhancement. In: British Machine Vision Conference. (2018)
30. Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). (2020) 1780–1789
31. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. ACM Trans. Graph. **35** (2016) 110:1–110:11
32. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV. (2016)
33. Zhao, J., Liu, L., Snoek, C.G.M., Han, J., Shao, L.: Pixel-level semantics guided image colorization. ArXiv **abs/1808.01597** (2018)
34. Su, J.W., Chu, H.K., Huang, J.B.: Instance-aware image colorization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2020)
35. Zhang, R., Zhu, J.Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time user-guided image colorization with learned deep priors. ACM Trans. Graph. **36** (2017) 119:1–119:11
36. Xiao, Y., Zhou, P., Zheng, Y.: Interactive deep colorization using simultaneous global and local inputs. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2019) 1887–1891
37. Dong, X., Li, W., Wang, X., Wang, Y.: Learning a deep convolutional network for colorization in monochrome-color dual-lens system. In: AAAI. (2019)
38. He, M., Chen, D., Liao, J., Sander, P.V., Yuan, L.: Deep exemplar-based colorization. ACM Transactions on Graphics (TOG) **37** (2018) 1 – 16

39. Zhang, B., He, M., Liao, J., Sander, P.V., Yuan, L., Bermak, A., Chen, D.: Deep exemplar-based video colorization. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 8044–8053
40. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 2472–2481
41. Liu, D., Wen, B., Fan, Y., Loy, C.C., Huang, T.S.: Non-local recurrent network for image restoration. In: NeurIPS. (2018)
42. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: ECCV. (2018)
43. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: CVPR. (2020)
44. Barbastathis, G., Ozcan, A., Situ, G.: On the use of deep learning for computational imaging. Optica **6** (2019) 921–943
45. Ironi, R., Cohen-Or, D., Lischinski, D.: Colorization by example. In: Rendering Techniques, Citeseer (2005) 201–210
46. Gupta, R.K., Chia, A.Y.S., Rajan, D., Ng, E.S., Zhiyong, H.: Image colorization using similar images. In: Proceedings of the 20th ACM international conference on Multimedia. (2012) 369–378
47. Liu, Z., Yeh, R.A., Tang, X., Liu, Y., Agarwala, A.: Video frame synthesis using deep voxel flow. In: Proceedings of International Conference on Computer Vision (ICCV). (2017)
48. Jiang, H., Sun, D., Jampani, V., Yang, M.H., Learned-Miller, E.G., Kautz, J.: Super slomo: High quality estimation of multiple intermediate frames for video interpolation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 9000–9008
49. Bao, W., Lai, W.S., Zhang, X., Gao, Z., Yang, M.H.: Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
50. Haris, M., Shakhnarovich, G., Ukita, N.: Space-time-aware multi-resolution video enhancement. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2020)
51. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 8934–8943
52. Gharbi, M., Chen, J., Barron, J.T., Hasinoff, S.W., Durand, F.: Deep bilateral learning for real-time image enhancement. ACM Transactions on Graphics (TOG) **36** (2017) 118
53. Chen, J., Paris, S., Durand, F.: Real-time edge-aware image processing with the bilateral grid. ACM Transactions on Graphics (TOG) **26** (2007) 103–es
54. Wikipedia contributors: Yuv — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=YUV&oldid=962998638` (2020) [Online; accessed 30-June-2020].
55. Schechner, Y.Y., Nayar, S.K., Belhumeur, P.N.: Multiplexing for optimal lighting. IEEE Transactions on Pattern Analysis and Machine Intelligence **29** (2007) 1339–1354
56. Jeon, H.G., Lee, J.Y., Im, S., Ha, H., Kweon, I.S.: Stereo matching with color and monochrome cameras in low-light conditions. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 4086–4094

57. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2015)
58. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In Jiang, X., Hornegger, J., Koch, R., eds.: Pattern Recognition, Cham, Springer International Publishing (2014) 31–42
59. Schwartz, E., Giryes, R., Bronstein, A.M.: Deepisp: Toward learning an end-to-end image processing pipeline. IEEE Transactions on Image Processing **28** (2019) 912–923
60. Welsh, T., Ashikhmin, M., Mueller, K.: Transferring color to greyscale images. In: Proceedings of the 29th annual conference on Computer graphics and interactive techniques. (2002) 277–280
61. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
62. Chen, C., Xiong, Z., Tian, X., Zha, Z.J., Wu, F.: Camera lens super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 1652–1660