RESEARCH ARTICLE

SRPS–deep-learning-based photometric stereo using superresolution images

Euijeong Song1, Seokjung Kim1, Seok Chung1,2,∗ and Minho Chang3,∗

1School of Mechanical Engineering, Korea University, Seoul 02841, Korea; 2KU-KIST Graduate School of Converging Science and Technology, Seoul 02841, Korea and 3Medit Corp., Seoul 02855, Korea

*Corresponding author. E-mail: sidchung@korea.ac.kr; minho@medit.co.kr

Abstract

This paper introduces a novel deep-learning-based photometric stereo method that uses superresolution (SR) images: SR photometric stereo. Recent deep-learning-based SR algorithms have yielded great results in terms of enlarging images without mosaic effects. Supposing that the SR algorithms successfully enhance the feature and colour information of original images, implementing SR images using the photometric stereo method facilitates the use of considerably more information on the object than existing photometric stereo methods. We built a novel deep-learning-based network for the photometric stereo technique to optimize the input–output of SR image inputs and normal map outputs. We tested our network using the most widely used benchmark dataset and obtained better results than existing photometric stereo methods.

Keywords: computer vision; photometric stereo; deep learning; convolutional neural network; image superresolution

1. Introduction

The photometric stereo (PS) method uses more than three images of an object captured by a single camera with varied directions of a light source and derives the surface normal data of the object. The PS method can derive detailed data on the surface texture and structure compared to other surface reconstruction methods, such as stereo matching and structured light. However, the main challenge of the PS method is that the reflection of light by the low-Lambertian material can interrupt the surface normal estimation. While measuring the surface normal data of an object with an ideal Lambertian surface, only three images with different light sources and simple matrix equations are sufficient to sustain accuracy. However, when the surface material of an object is non-Lambertian, such as glossy plastic or gold, and the intensity of light reflected from the shiny object is not proportional to its slanted rate, the traditional PS methods can generate large measurement errors. To overcome these obstacles, optimization-based PS methods have been widely studied; this requires a large number of images and high computational costs. The quality of the resulting normal maps derived using these methods is not stable. The method proposed by Aldrin et al. (2008) exhibits great performance for semi-Lambertian objects. However, it is not consistent for various semi-Lambertian objects and is especially bad for metallic surfaces.

Recently, deep-learning methods have been applied in various computer vision divisions, such as stereo matching and PS. Santo et al. (2017) developed a deep PS network (DPSN), wherein, for the first time, the PS adopted a deep neural network, employing dense neural layers and several dropout layers to its network to reduce errors caused by shadows. In the PS-convolutional neural network (CNN) method (Ikehata, 2018), an observation map was derived using every pixel in an image, and the observation map was placed in a CNN-based network to calculate the surface normal vector corresponding to each pixel, which yielded good results for non-Lambertian objects. Zheng et al. (2019) developed the idea by Ikehata (2018) and introduced SPLINE-Net by applying a two-step CNN-based network.
Further, they invented a symmetric/asymmetric loss function to enhance the training results and reduce the number of images required. However, Zheng et al. (2019) reported that the single-pixel processing PS method can provide inaccurate normal maps even for Lambertian objects. PS-FCN (Chen et al., 2018) utilizes a large amount of information from neighboring pixels and is known as the most stable method for existing deep-learning-based PS methods; hence, it is considered to be the best; however, the surface normal map data derived by the PS-FCN lack high-frequency information.

Meanwhile, deep-learning methods are also utilized in image restoration areas for purposes such as noise reduction and SR. Single-image super-resolution (SISR) problem is considered very difficult to solve. However, deep-learning-based methods have demonstrated that it is possible to derive a high-quality SR image. We propose that if the SR method can successfully enlarge the original images for the PS method with slight information corruption, it is possible for the PS method to use more data from the input images. We implemented the residual channel attention network (RCAN) algorithm (Zhang et al., 2018) to enlarge the images by two times and applied our novel CNN-based network to derive an original-sized normal map. Our method yielded better results than other existing PS methods, especially in eliminating low-frequency errors while preserving the high-frequency information of the object.

2. Previous Works

The first PS was proposed by RJ Woodham (1980) and used three light sources with a calibrated direction vector and a single orthographic camera. With three direction vectors for each light source, three images are taken with each light source on, and the surface orientation vector of a surface point is calculated using the three intensity values of the image pixel corresponding to the surface point. The calibration method introduced by Hayakawa (1994) measures the reflectance data of a known shape. Because a sphere is straightforward while determining the gradient of the shape, they chose a sphere as the reference shape, which is still being adopted in current light calibration. However, these methods cannot derive accurate surface normal vectors for a non-Lambertian surface because the equations presented by Woodham (1980) and Hayakawa (1994) do not fit the reflection characteristic of the non-Lambertian surface. To solve this problem, many papers (Iwahori et al., 1995; Wu et al., 2010; Shi et al., 2013; Ikehata & Aizawa, 2014) have proposed an optimization-based solution. These studies could derive an improved normal map compared to the conventional rule-based PS method; however, they still required more than 30 images for accurate results, and the assumptions they made did not perfectly fit materials in the real world, thereby making the results unstable. Thus, further improvement was required. In 2019, Chen (2019) devised a method that estimates the reflectance, smoothness, and normal of the surface using a non-deep-learning method in a microfacet-based reflectance model, which could replace the existing surface reflectance estimation model effectively, to obtain results close to those of deep-learning methods.

Recently, studies on deep-learning-based PS have been conducted with improved results. The first deep-learning-based PS was established by Santo et al. (2017) with 96 images per object. Because the input of the network is a group of pixels of images having the same coordinates, it is possible to train a network with a relatively small dataset. Ikehata et al. (2018) introduced a PS-CNN that utilizes an observation map as the input of its network. Zheng et al. (2019) improved the observation map, applied symmetric/asymmetric cost during training, and succeeded in obtaining good results with a relatively small number of images. Networks introduced in the PS-CNN and SPLINE-Net use one pixel per image to derive one pixel of the normal map; hence, they require a small computational cost. Li et al. (2019) fixed the range of the angle of the light position of the training dataset and effectively reduced the number of images required and found the optimal group of light sources.

However, Zheng et al. (2019) reported that the networks failed to calculate precise normal maps when an object had a Lambertian surface. While deriving the surface normal vector of a pixel, the DPSN, PS-FCN, and SPLINE-Net do not refer to any information from neighbouring pixels; therefore, the normal vector is not properly derived when the light information of the pixel is insufficient (e.g. a small number of images of a textureless non-Lambertian object).

Chen et al. (2018) proposed a network named PS-FCN that max-pools the feature maps from each image and de-convolutes the fused feature map to derive the final normal map. They then improved their method by normalizing the input images, which eliminated the defects caused by the albedo difference (Chen et al., 2020). These methods use N-shared encoder networks and one decoder network while using N inputs and, therefore, require a large amount of resources. PS-FCN and PS-FCN+N use a large number of neighbouring pixels per surface normal vector; consequently, they output normal map results with less low-frequency errors than any other PS method but lose high-frequency information of the surface normal. The overall summary of the deep-learning-based PS method has been presented by Zheng et al. (2020). According to the authors, PS-FCN+N is considered to be the state-of-the-art method with respect to the average accuracy and stability while using a small number of images.

Meanwhile, image SR is a sub-area of image restoration methods that restores the information loss caused by image size reduction. Before the growth of deep learning, image SR methods analysed the image using continuous Fourier transform (CFT), discrete Fourier transform (DFT), and multiple continuous images (Liakakhusis & Ananthashayana, 2008; Zheng et al., 2010). The point of SR methods that use CFT/DFT analysis is that the image registration can help interpolate low-resolution images onto a high-resolution image grid, and CFT/DFT analysis can efficiently extract suitable information for image registration. CFT/DFT can be replaced by a wavelet transform (Iwahori et al., 1995; Shi et al., 2016). Non-deep-learning-based SISR assumes that the image can be effectively restored if the edge information is accurately determined and a proper interpolation method is applied. Various non-deep-learning SISR methods have been introduced previously (Patel et al., 2011; Makwana & Mehta, 2013).

The deep-learning-based SR method was first developed by Dong et al. (2014) and exhibited extraordinary performance compared to non-deep-learning methods. The first deep-learning-based SISR, which used four convolutional layers to upscale the images by factors of 2, 3, and 4, was introduced by these authors. Among the various deep-learning-based SISR methods (Dong et al., 2014; Kim et al., 2016; Lim et al., 2017; Zhang et al., 2018), the RCAN introduced by Zhang et al. (2018) applied a residual-in-residual network to obtain high-quality SR images and was considered the most precise SR method until 2019, according to a review by Wang et al. (2020). After the RCAN, several methods (Jiang et al., 2020; Jin et al., 2020; Mei et al., 2020) were developed by
adjusting the hierarchical/multiple branch structures, and these exhibited better performance than the RCAN, especially when images were magnified by more than four times. While adjusting the $2\times$ SR on images, the PSNR value of the results (Jiang et al., 2020; Jin et al., 2020; Mei et al., 2020) did not exhibit a significant difference compared to the results obtained by Wang et al. (2020). Because we applied two-time SR on the input image, the results of the methods did not show a significant difference. As we wanted to adjust the $2\times$ SR, we implemented RCAN on the datasets and designed the architecture of a network suitable for SR images to obtain the desired results.

3. Architecture of the SRPS Network

Our method is based on the PS-FCN$^{13}$, which extracts N feature maps from N input images and merges the feature maps to derive one surface normal map. To enhance the performance of deep-learning-based networks, we improved the following aspects:

1) Using SR images without further pre-processing does not allow the network to calculate an accurate surface normal map. We established pre-processing to put an SR image into our network.

2) We redesigned the feature map network and cost function to use SR images as inputs. We found that using SR images and existing feature extracting networks without change does not enhance the details; therefore, we designed a proper network structure. Further, we designed the cost function to enhance the training by considering the normal vector and gradient of the normal vector.

We will describe the designed network structure and image pre-processing, and demonstrate how these factors enhanced the surface normal map calculation results. The overall flow of the network is presented in Fig. 1.

### 3.1. Image pre-processing

#### 3.1.1. Image normalization

The image pre-processing involves three steps: image normalization, image SR, and image–light vector concatenation. In the image normalization step, we adjusted equation (1) to each image that participates in the PS.

$$I_{kN} = \frac{I_k}{\sqrt{I_1^2 + I_2^2 + \ldots + I_N^2}}$$

Here, $I_k$ is the intensity value of a pixel in the kth image, N is the number of images that participate in the PS, and $k \in \{1, 2, \ldots, N\}$. $I_{kN}$ is the intensity of normalized image $I_{kn}$. The normalization process was performed independently for each RGB colour channel.

If the albedo of an object is very small or large, the intensity difference between the participating images reduces regardless of the light direction, which makes it difficult to calculate the surface normal vector accurately. Normalizing images solves this problem by eliminating the effects of albedo. Although the oversaturation of light is not perfectly reduced by normalization, it helps in the overall process of the PS. An example of the normalization results is presented in Fig. 2. Focusing on the black stripe texture of the object, it can be observed that the black texture is removed from the normalized image.

#### 3.1.2. Image SR

After the normalization process, the input images were enlarged using the SR method. The existing SR method multiplies the resolution of the input image by scaling factors of 2, 3, and 4. According to Zhang et al. (2018), the PSNR rate of SR images is best for two-time multiplied images, which means that SR images enlarged by a scaling factor larger than 2 contain more errors. We empirically concluded that it is not necessary to enlarge images more than two times and used $2\times$ SR images for our network.
We designed our network to obtain \((2h \times 2w)\)-sized images as inputs and generate an \((h \times w)\)-sized surface normal map as the output; thus, only the input images are SR. Here, \(h\) and \(w\) are the height and width of the output normal map, respectively. The result obtained through the SR step is presented in Fig. 3. Comparing the hair and ear parts of the images, it can be observed that the SR image contains more detailed information regarding the object.

To ensure that we have correctly adjusted the SR method introduced by Zhang et al. (2018), we tested our implemented
Table 1: Comparison of SR effect between original and normalized images.

|                  | Original image | Normalized image |
|------------------|----------------|------------------|
|                  | Before SR      | After SR         | Δ (%)  | Before SR      | After SR         | Δ (%)  |
| Average          | 0.034          | 0.067            | 97.1   | 0.098          | 0.146            | 49.0   |
| Std. dev.        | 0.032          | 0.014            | 56.3   | 0.051          | 0.042            | 17.6   |

\[ v = \{v_{x1}, v_{y1}, v_{z1}\} \subseteq \{0.0606, 0.4484, 0.8918\} \]

Figure 5: Description of light vector tile generation and concatenation process.

\[ f = [64, 128, 256, 512, 256] \]

Figure 6: Structure of the FEN.
Deep learning photometric stereo using superresolution images

Figure 7: Examples of feature maps. We chose four feature maps from 256 channels, which seems the most meaningful.

Figure 8: Max pooling layer description.

algorithm on the FlyingThings3D dataset (Jin et al., 2020), which is the most widely used dataset for stereo matching. We randomly selected 1000 images from the dataset, resized the original image by a scale factor of 0.5, and adjusted the SR to the image. According to Zhang et al. (2018), the PSNR value is used as the criterion for the performance of the algorithm. We calculated the PSNR value for each pair of images and confirmed that the calculated PSNR values varied in the range of 30.5–33.1 dB, which indicates that the algorithm was well implemented. Figure 4 presents an example of the implementation of the SR algorithm. The PSNR value of the pair in Fig. 4 is 31.65 dB.

Also, we calculated the PSNR value for image pair of normalized images in the DiLiGenT dataset (Shi et al., 2016), which will be used as the test dataset of our experiment. We normalized the original images of the dataset (Image A), resized the normalized images by 0.5× rate, and adjusted SR by 2.0× rate (Image B). We then derived the PSNR value of the Image A and Image B. The PSNR value of the pairs varied in the range of 28.6–30.7 dB. Although the SR network we implemented did not use normalized images as its training set, the network showed that it also worked properly for normalized images.

During the experiment, we found that the error caused by SR significantly affects the result, and the sequence of pre-processing is important to compensate for the error. SR effectively preserves the edge and detailed information regarding the original image but slightly loses light consistency. We found that adjusting the SR image on normalized images can preserve more lighting consistency than adjusting that on non-normalized images.

To demonstrate the intensity distortion caused by SR, we randomly selected 50 images from the DiLiGenT dataset (Shi et al., 2016) and derived the average and standard deviation (std. dev.) of each value. Then, we adjusted the SR on each image and compared the average and std. dev. difference between the original and SR images. Further, we normalized the images and repeated the process. Table 1 lists the average values of the 50 trials for each case. It should be noted that adjusting the SR on the non-normalized image significantly lowers the std. dev. value, whereas that of the normalized image is not lowered. These results explain why networks trained using SR images without image normalization fail to calculate surface normal maps accurately, whereas those using image normalization are successful.
3.1.3. Light vector–image concatenation

The final step in image pre-processing is the light vector tile generation and concatenation. The feature extraction step takes $N \times (2w \times 2h \times 6)$ arrays as inputs, where $N$ is the number of images used in the PS. Suppose that the input images are SR images and have a size of $(2w \times 2h \times 3)$. Because we want our PS to utilize the image as well as light vector information, we edited the input image by referring to Chen et al. (2018). To utilize the light direction vector efficiently, we created a tile matrix with a light vector. We duplicated the light direction vector and created a tile matrix with a size of $(2h \times 2w \times 3)$, similar to an SR input image. By concatenating the input image and tile matrix, the size of the input becomes $(2h \times 2w \times 6)$. This pre-processing of the input image simplifies the addition of inputs into our network, and the structure of the feature extraction network (FEN) is simple. The final step in image pre-processing is depicted in Fig. 5.

3.2. FEN

The pre-processed images were placed in the FEN. The FEN structure is composed of several CNN layers. The entire structure of the FEN is shown in Fig. 6. We stacked 11 CNN layers to set the output of FEN as $(h/2 \times w/2 \times 256)$, referring to the
Figure 11: Example images from the benchmark dataset.

Table 1: Error map of L2 method result

|                  | L2 group 1 | L2 group 2 | L2 group 3 | L2 group 4 |
|------------------|------------|------------|------------|------------|
| L2(rad⁻¹)        | 13.97      | 6.83       | 6.81       | 6.45       |

Table 2: Mean of Gradient of GT Normal map = Detail

|                  | MOG group 1 | MOG group 2 | MOG group 3 |
|------------------|-------------|-------------|-------------|
| MoG(rad)         | 0.22        | 0.31        | 0.34        |

Figure 12: Test dataset grouping by two criteria. Upper: Inverse of L2⁻¹. Lower: MoG.

Figure 13: Local areas chosen for Local MAE analysis.
structure of the PS-FCN. We then added two skip connections between layers 5 and 8 and layers 3 and 10. We set the size of the output as similar to the final normal map to prevent the information loss that may potentially occur via deconvolution in the final step of the network.

According to ResNet (He et al., 2016) and U-Net (Ronneberger et al., 2015), building skip connections between layers helps us to detect fine features from images, making it widely applicable in other deep-learning-based computer vision areas. We expected the skip connection to extract detailed feature maps from the input images. If the number of images used is N, the FEN is called N times, and N feature maps are extracted and passed to the next step. Additionally, Fig. 7 presents an example of the resulting feature maps. Some layers of the feature map contain meaningful information, such as the edges of the object (channel 24) or shadow part caused by the light source (channel 131).

3.3. Merge and calculation network (MnC)

After extracting N feature maps, they are merged and passed to the MnC to calculate the normal map. We adopted a max pooling layer to fuse the feature maps from the FEN. Because pooling layers are not weight based, the order of the light vector does not affect the results. The role of max pooling is illustrated in Fig. 8.

The normal map calculation network is composed of seven 2D convolutional layers, two skip residual shortcuts, and an L2-normalization layer. Skip connections were also applied to layers 1, 3, 4, and 7. In the ablation section, it is shown that adding a skip connection to the network is essential because the stacked convolutional layer cannot properly utilize the information of the SR image. The structure of the MnC is shown in Fig. 9.

3.4. Loss function for learning

Because the measurement of the output is achieved by calculating the cosine similarity function, it is efficient to use the function as the loss function of our network as well. The cosine similarity loss function L is as follows:

\[ L = \frac{1}{h \times w} \sum_{i,j} \left( 1 - n_{ij} \cdot n'_{ij} \right). \]
Deep learning photometric stereo using superresolution images

Figure 16: Sample images from the synthetic dataset: (a) alien, (b) house, and (c) octopus.

Figure 17: Resulting surface normal map for the synthetic dataset calculated by our method.

Table 2: MAE comparison for PS-FCN+N and our method on the synthetic datasets.

|        | Alien1 | Alien2 | House1 | House2 | Octopus1 | Octopus2 |
|--------|--------|--------|--------|--------|----------|----------|
| PS-FCN+N | 6.75   | 8.11   | 6.97   | 8.40   | 8.54     | 9.23     |
| Ours    | 4.16   | 5.28   | 5.57   | 5.30   | 7.21     | 7.02     |

Here, $n_i$ and $n'_i$ denote the calculated surface normal vector and ground truth vector, respectively. The maximum value of $n_i \cdot n'_i$ is 1 when $n_i$ and $n'_i$ are equal; thus, the cosine similarity loss function is adequate for use.

Since the loss function above considers the mean value of angular difference only, it does not guarantee the detail information of surface, especially when the algorithm utilizes lots of neighbouring information; thus, it causes blurring effect. The
Figure 18: Comparison of the results obtained using two networks: (a) ours and (b) PS-FCN+\textsuperscript{vN}.

Table 3: Error value comparison between $N_o$ and $N_r$. Note that the average values of each of the normal map groups are almost same.

|     | Ball | Bear | Cat | Pot1 | Pot2 | Buddha | Goblet | Reading | Cow | Harvest | Average |
|-----|------|------|-----|------|------|--------|--------|---------|-----|---------|---------|
| $N_o$ | 5.01 | 6.09 | 7.49 | 9.19 | 10.13 | 8.02   | 13.48  | 12.82   | 10.21| 18.47   | 10.09   |
| $N_r$ | 2.43 | 4.85 | 7.62 | 8.05 | 10.28 | 13.72  | 14.40  | 12.05   | 9.54 | **18.04** | 10.10   |

Figure 19: Derived $N_o$ and $N_r$. (a): $N_o$. (b): $N_r$. 
Table 4: Comparison of normal map calculation results on the benchmark dataset.

| Number of Images Used | PS-FCN, PS-FCN\textsuperscript{IN}, CNN-PS, Ours: 16 | SPLINE-Net: 10 | Others: 96 |
|-----------------------|-----------------------------------------------|-----------------|-------------|
| Grouped by L2 Error   | PS-FCN, PS-FCN\textsuperscript{IN}, CNN-PS, Ours: 16 | SPLINE-Net: 10 | Others: 96 |
| Grouped by MoG        | PS-FCN, PS-FCN\textsuperscript{IN}, CNN-PS, Ours: 16 | SPLINE-Net: 10 | Others: 96 |

Here, $\Delta n_x$ and $\Delta n'_x$ denote the $x$-direction gradient of the derived normal vectors and ground truth normal vectors, respectively; $\Delta n_y$ and $\Delta n'_y$ are the $y$-direction gradients of derived normal vectors and the ground truth normal vectors, respectively.

Using $L_C$ only can cause the global error since angular error is not considered in $L_C$. The final loss function, $L_f$, is established.
Table 5: MAE-LG comparison of results for our network and existing networks.

| MAE-LG    | Bear | Cat | Pot1 | Pot2 | Buddha | Goblet | Reading | Cow | Harvest | Average |
|-----------|------|-----|------|------|--------|--------|---------|-----|---------|---------|
| PS-FCN    | 1.00 | 1.31| 2.12 | 2.92 | 5.54   | 4.94   | 7.82    | 2.08| 7.12    | 3.85    |
| PS-FCN-N^2| 0.47 | 1.14| 1.61 | 1.79 | 2.71   | 3.07   | 3.61    | 1.94| 3.85    | 2.47    |
| SRPS(our) | 0.29 | 0.65| 1.29 | 1.40 | 2.37   | 2.57   | 3.01    | 1.16| 4.93    | 1.96    |

Table 6: Local MAE comparison.

| Local MAE | PS-FCN | PS-FCN-N^2 | Ours |
|-----------|--------|------------|------|
| Local 1   | 5.388  | 4.812      | 3.378|
| Local 2   | 12.67  | 5.902      | 3.305|
| Local 3   | 12.66  | 8.174      | 5.526|

We empirically chose $\alpha$ as 0.4, which showed the best result among the experiments done with various values between 0.1 and 0.9.

4. Experimental Setting, Dataset, and Result Criteria

In this section, we will describe the dataset, hardware, software set-up, and criteria for performance analysis.

Table 7: Object MAE comparison for ablation study.

| Object MAE | Bear | Cat | Pot1 | Pot2 | Buddha | Goblet | Reading | Cow | Harvest | Average |
|------------|------|-----|------|------|--------|--------|---------|-----|---------|---------|
| No SR Img. | 4.78 | 6.68| 7.22 | 8.38 | 10.48  | 10.37  | 11.70   | 10.88| 15.97   | 9.72    |
| No skip-Con.| 4.71 | 7.45| 7.20 | 8.38 | 10.63  | 10.57  | 11.52   | 10.45| 17.89   | 9.87    |
| SRPS(our)   | 3.37 | 4.72| 5.24 | 6.53 | 8.69   | 8.87   | 10.14   | 7.72 | 15.59   | 7.87    |

Table 8: MAE-LG comparison for the ablation study.

| MAE-LG | Bear | Cat | Pot1 | Pot2 | Buddha | Goblet | Reading | Cow | Harvest | Average |
|--------|------|-----|------|------|--------|--------|---------|-----|---------|---------|
| No SR Img. | 0.51 | 0.97| 1.44 | 1.65 | 2.74   | 2.88   | 3.08    | 1.56| 5.23    | 2.23    |
| No skip-Con.| 0.93 | 1.92| 2.22 | 2.77 | 3.51   | 4.13   | 4.12    | 2.31| 7.96    | 3.25    |
| SRPS(our) | 0.29 | 0.65| 1.29 | 1.40 | 2.37   | 2.57   | 3.01    | 1.16| 4.93    | 1.96    |

4.1. Training dataset

We used the Sculpture PS Dataset, which was implemented by Chen et al. (2018). The Sculpture PS Dataset employs eight 3D models with a complicated shape. Apparently, 3700 views per 3D model were applied, and two randomly different surface materials were rendered on the model. Each viewpoint has 64 random light conditions; hence, the total number of samples was $29 \times 64 \times 2 = 292$, and the total number of images was $59 \times 292 \times 64 = 3794 \times 688$. The original size of the dataset images was $128 \times 128$, and we adjusted the RCAN SR algorithm for each image to obtain a size of $256 \times 256$. We applied pre-trained RCAN to input images of the dataset only and did not apply it to the normal map of the dataset, which was used as the ground truth. We cropped patches from each image by $96 \times 96$ and $48 \times 48$ from the normal map. Example images of the dataset used are presented in Fig. 10.

To train our network, we implemented the Keras framework (Keras, 2015). We randomly split the dataset into training and validation sets at a ratio of 99:1. The initial learning rate was set to 0.001 and was reduced by half per five epochs. We trained our network with a batch size of 32 for 50 epochs. The total number of weights for our network was 10 275 587; hence, the entire training process took approximately 3 days. We observed that...
the final loss value oscillates near 0.14, and the validation loss significantly oscillates; therefore, we did not consider the validation loss as the dominant determinant for obtaining good training results. The loss oscillation may have been caused by the effect of the black part of the dataset images. We employed the Adam optimizer (Kingma & Jimmy, 2014) to train with the default parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$). We used Nvidia Tesla V100 with 32 GB of VRAM and 128 GB of RAM for training. We observed that the minimum RAM requirement for training our network was 32 GB.

4.2. Test dataset

The most widely used benchmark dataset for the PS method is the DiLiGenT dataset (Shi et al., 2016), which contains 10 objects with random and composite materials. Each object has 96 images with different light directions. Examples of images and the ground truth are presented in Fig. 11. The input images are also superresolutionized by a scale factor of 2 before being put into our network. We chose 16 as the number of images to use, $N$, and compared the resulting surface normal map with those obtained in other methods.

To show the contribution of our research more clearly, we grouped the test datasets by two criteria. The first criterion is the inverse of the L2 (Woodham, 1980) error value of each object ($L_2^{-1}$). The L2 error becomes zero when the object is completely Lambertian and shadowless; therefore, we used $L_2^{-1}$ as a criterion for the Lambertian-ness of the object. The second criterion is the mean of gradient (MoG), which is the mean of the gradient of the ground-truth normal. If the MoG value is large, the object has a complex shape. Figure 12 shows the grouping of the objects using the two criteria.

4.3. Error evaluation criteria

To evaluate the performance of our network, we devised the following error criteria:

- Object MAE: Mean angular error value of the derived normal map of an object.
- Local MAE: Mean angular error value of the derived normal map of the selected area. We chose three local areas from Buddha and Harvest, which have the most complex shapes. The three local areas are illustrated in Fig. 13.
- MAE-LG (MAE-large gradient): Mean angular error value of the derived normal map of the area with a large gradient on the ground-truth normal map.

A large gradient indicates that the geometry of the area is complex. Therefore, comparing the MAE of the large-gradient area can help evaluate the performance of each method on a detailed surface. We filtered the pixels with a gradient smaller than the median gradient value of the object. Figure 14 shows the filtered normal maps for deriving the MAE-LG. The ball object is neglected in this step because we deduced that comparing MAE-LG on a completely smooth object is meaningless.

4.4. Ablation study

We trained two networks for the ablation study. The difference between our network and each network is as follows:

1. Net 1: Net 1 uses bicubic interpolated images as input to check the effect of the SR image input.
2. Net 2: Net 2 has no skip connection in the network, as in PS-FCN.

The structure of each network is illustrated in Fig. 15. We tested each network on the test dataset and derived the MAE and MAE-LG values for the ablation study.

5. Test Results and Discussion

5.1. Synthetic dataset test

We created six synthetic datasets using three complicated objects: alien, house, and octopus. We randomly picked six surface materials. The sample images are shown in Fig. 16. Each object has $11 \times 11$ regularly distributed light positions, and we randomly selected 16 positions to obtain a surface normal map. The derived surface normal map is shown in Fig. 17, and the MAE errors of each normal map are listed in Table 2. Our methods outperformed the PS-FCN on every object in the synthetic dataset. We averaged 50 trial results for each object. You can see the detail of a surface normal map in Fig. 18.

5.2. SR image quality test

In this chapter, we validate how much SR degrades the result of PS. Since the original DiLiGenT dataset does not have (2×)-sized image, we made two normal maps for error comparison as follows:

- Normal map from original normalized image ($N_o$): We derived a normal map $N_o$ using normalized DiLiGenT dataset images. To obtain the result, we trained a network SRPS', which has the same structure with SRPS but is trained using normalized images of training dataset without SR, so that it outputs (0.5×)-sized normal maps $N_o$ as a result.
- Normal map from resized-SR normalized images ($N_r$): We resized the normalized DiLiGenT dataset images by 0.5× rate. Then, we adjusted 2.0× SR to the images to obtain resized-SR images, which has the same size with the original images. We utilized these images as input of SRPS and derived (0.5×)-sized normal maps ($N_r$). By comparing the mean angular error of $N_o$ and $N_r$, we can measure the degradation effect of SR on PS.

Table 9: Local MAE comparison results for the ablation study.

|          | Net 1 | Ours | Net 2 | Ours |
|----------|-------|------|-------|------|
| Local 1 MAE | 3.810 | 3.378 | 4.601 | 3.378 |
| Local 2 MAE | 3.673 | 3.305 | 3.582 | 3.305 |
| Local 3 MAE | 6.968 | 5.526 | 7.184 | 5.526 |
Figure 21: Comparison of results for the local part of the benchmark dataset. From left to right: ground truth, PS-FCN result, and SRPS (ours) result.

Table 10: MAE result when using $L$ and $L_f$ as loss functions for training.

| MAE  | Ball | Bear | Cat  | Pot1 | Pot2 | Buddha | Goblet | Reading | Cow | Harvest | Average |
|------|------|------|------|------|------|--------|--------|---------|-----|---------|---------|
| $L$  | 3.27 | 6.09 | 6.12 | 6.20 | 7.07 | 10.02  | 9.25   | 11.32   | 7.63| 17.00   | 8.34    |
| $L_f$ | 2.83 | 3.37 | 4.72 | 5.24 | 6.53 | 8.69   | 8.87   | 10.14   | 7.72| 15.59   | 7.37    |

Table 11: MAE-LG result when using $L$ and $L_f$ as loss functions for training.

| MAE-LG | Bear | Cat | Pot1 | Pot2 | Buddha | Goblet | Reading | Cow | Harvest | Average |
|--------|------|-----|------|------|--------|--------|---------|-----|---------|---------|
| $L$   | 0.52 | 0.98| 1.80 | 2.06 | 3.32   | 3.88   | 3.23    | 1.57| 6.23    | 2.36    |
| $L_f$ | 0.29 | 0.65| 1.29 | 1.40 | 2.37   | 2.57   | 3.01    | 1.16| 4.93    | 1.96    |

The result normal maps and error comparison are shown below. Table 3 shows the mean angular value of $N_o$ and $N_r$. Since there is small difference between two values and superiority cannot be determined, we concluded that SR does not degrade the performance of our method. Fig. 19 shows the samples of $N_o$ and $N_r$. From this result, we inferred that our network effectively compensates the error of image by the degradation of SR.
Deep learning photometric stereo using superresolution images

Figure 22: Normal map derivation results for light-stage data gallery dataset. Left: Our method. Right: PS-FCN.

5.3. Overall comparison of benchmark dataset

Table 4 presents the results of the comparison. We selected 16 random images for each trial, and the resulting values of our methods were the average of 30 trials per object. Our method exhibits the best performance among the PS methods for every object in the dataset, except for the Ball and Harvest datasets. Examples of the results of the normal and error maps are shown in Fig. 20.

Table 5 presents the MAE-LG errors for our network as well as the PS-FCN and PS-FCN\(^{N}\). Our network exhibited the best performance for every object in the dataset. Table 6 presents the local MAE results of our network, along with those for the PS-FCN and PS-FCN\(^{N}\). Our network exhibited the best performance for every object in the dataset. The MAE-LG result shown in Tables 5 and 6 shows that our method has superiority on calculating the surface normal vector of a complex object. Additionally, the MAE-LG results reveal that our method achieves better results for all objects, especially for MoG group 1.

Tables 7 and 8 present the ablation study results with respect to object MAE and MAE-LG. Remarkably, using bicubic images as input instead of SR images also yielded good performance on the MAE-LG comparison, but it showed relatively low performance on the object MAE. It can be deduced that bicubic interpolation can preserve the detailed information of an object but exhibits inferior performance in preserving the material characteristics of the surface.

Table 9 presents the ablation study results obtained using the local MAE results, and Fig. 21 shows the normal map calculation results for the local MAE comparison between the PS-FCN\(^{N}\) and our method. It can be observed that detailed information is well preserved in the results from our methods, such as the teeth of Buddha as well as the hairband jewel and cloth ornaments of Harvest.

Tables 10 and 11 present the ablation study that compares results (MAE error/MAE-LG error) when our network is trained with different loss functions \(L\) and \(L_f\). The result shows that using \(L_f\) derives superior result for all objects in test dataset than using only \(L\).
6. Conclusion and Future Work

In this paper, we proposed a new PS method that utilizes an SR image; this method is referred to as SRPS. Our SRPS method can derive more precise and accurate surface normal map data using a relatively small number of images. Additionally, our network efficiently reduces the high-frequency error caused by several factors compared to other methods. These results indicate that using an SR image on a computer vision/image analysis area is considerably effective, because the amount of applicable information drastically increases. We will identify other areas such as stereo vision or image segmentation, where SR is also applicable.

Because our network uses many neighbouring pixels per surface normal vector, the detailed information still needs to be improved. Moreover, our network has a larger network size than other deep-learning-based networks, which increases further when the RCAN network is considered simultaneously (the total number of weights is approximately 13 million). This can be a critical disadvantage for practical use; therefore, we will find ways to optimize the number of weights in our future work. Additionally, our network is available for only calibrated light environments; consequently, we are planning to reinforce our network for implementation in uncalibrated light environments.

Further, we tested our network on light-stage data gallery datasets (Einarsson et al., 2006) but obtained poor results for all the objects. This is because the object was so large that the light could not cover its entire area, thus causing a significant error while using a small number of images. In Fig. 22, it can be observed that the resulting normal map for the Knight-Standing and Helmet is defective and blurred, compared to the result for the PS-FCN, which used approximately 140 images. This indicates that every light source used for our method should cover the entire part of the object to be measured to make our network function effectively.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSST: 2020R1A2B5B03002005). Also, this work was a part of the ‘HPC Support’ Project, supported by the ‘Ministry of Science and ICT’ and the National IT Industry Promotion Agency. We sincerely appreciate this project for providing a high-spec research environment.

Conflict of interest statement

None declared.

References

Alldrin, N., Zickler, T., & Kriegman, D. (2008). Photometric stereo with non-parametric and spatially-varying reflectance. In 2008 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1–8).

Chen, G., Han, K., & Wong, K.-Y. K. (2018). PS-FCN: A flexible learning framework for photometric stereo. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 3–18).

Chen, G., Han, K., Shi, B., Matsushita, Y., & Wong, K.-Y. K. (2020). Deep photometric stereo for non-Lambertian surfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2020.3005397.

Chen, L., Zheng, Y., Shi, B., Subpa-asu, A., & Sato, A. (2019). A microfacet-based model for photometric stereo with general isotropic reflectance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43, 48–61.

Dong, C., Loy, C. C., He, K., & Tang, X. (2014). Learning a deep convolutional network for image super-resolution. In European Conference on Computer Vision (pp. 184–199).

Einarsson, P., Chabert, C. F., Jones, A., Ma, W. C., Lamond, B., Hawkins, T., Bolas, M., Sylwan, S., & Debevec, P. (2006). Relighting human locomotion with flowed reflectance fields. In Proceedings of the Eurographics Conference on Rendering Techniques (pp. 183–194).

Hayakawa, H. (1994). Photometric stereo under a light source with arbitrary motion. Journal of the Optical Society of America A, 11, 3079–3089.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 770–778).

Ikehata, S. (2018). CNN-PS: CNN-based photometric stereo for general non-convex surfaces. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 3–18).

Ikehata, S., & Aizawa, K. (2014). Photometric stereo using constrained bivariate regression for general isotropic surfaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2179–2186).

Iwahori, Y., Woodham, R. J., & Bagheri, A. (1995). Principal components analysis and neural network implementation of photometric stereo. In Proceedings of the Workshop on Physics-Based Modeling in Computer Vision (p. 117).

Jiang, K., Wang, Z., Yi, P., & Jiang, J. (2020). Hierarchical dense recursive network for image super-resolution. Pattern Recognition, 107, 107475.

Jin, J., Hou, J., Chen, J., & Kwong, S. (2020). Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2260–2269).

Keras. Deep learning for humans. (2015). http://keras.io.

Kim, J., Kwon Lee, J., & Mu Lee, K. (2016). Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1646–1654).

Kingma, D. P., & Jimmy, B. A. (2014). Adam: A method for stochastic optimization. arXiv preprint: arXiv:1412.6980.

Li, J., Robles-Kelly, A., You, S., & Matsushita, Y. (2019). Learning to minimize photometric stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7568–7576).

Lim, B., Son, S., Kim, H., Nah, S., & Mu Lee, K. (2017). Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 136–144).

Liyakathunisa, , & Ananthashayana, V. K. (2008). Superresolution blind reconstruction of low-resolution images using wavelets based fusion. World Academy of Science, Engineering and Technology, 40, 177–181.

Makwana, R. R., & Mehta, N. D. (2013). Survey on single image super resolution techniques. IOSR Journal of Electronics and Communication Engineering (IOSR-JECE), 5, 23–33.

Mei, Y., Fan, Y., Zhou, Y., Huang, L., Huang, T. S., & Shi, H. (2020). Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5690–5699).
Patel, V., Modi, C. K., Paunwala, C. N., & Patnaik, S. (2011). Hybrid approach for single image super resolution using ISEF and IBP. In 2011 International Conference on Communication Systems and Network Technologies (pp. 495–499).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 234–241).

Santo, H., Samejima, M., Sugano, Y., Shi, B., & Matsushita, Y. (2017). Deep photometric stereo network. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 501–509).

Shi, B., Tan, P., Matsushita, Y., & Ikeuchi, K. (2013). Bipolynomial modeling of low-frequency reflectances. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36, 1078–1091.

Shi, B., Wu, Z., Mo, Z., & Duan, D. (2016). A benchmark dataset and evaluation for non-Lambertian and uncalibrated photometric stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3707–3716).

Woodham, R. J. (1980). Photometric method for determining surface orientation from multiple images. Optical Engineering, 19, 191139.

Wang, Z., Chen, J., & Hoi, S. C. H. (2020). Deep learning for image super-resolution: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2020.2982166.

Wu, L., Ganesh, A., Shi, B., Matsushita, Y., Wang, Y., & Ma, Y. (2010). Robust photometric stereo via low-rank matrix completion and recovery. In Asian Conference on Computer Vision (pp. 703–717).

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018). Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 286–301).

Zheng, H., Bouzerdoum, A., & Phung, S. L. (2010). Wavelet-based nonlocal-means super resolution for video sequences. In Proceedings of 2010 IEEE 17th International Conference on Image Processing (pp. 2817–2820).

Zheng, Q., Jia, Y., Shi, B., Jiang, X., Duan, L.-Y., & Kot, A. C. (2019). SPLINE-Net: Sparse photometric stereo through lighting interpolation and normal estimation networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 8549–8558).

Zheng, Q., Shi, B., & Pan, G. (2020). Summary study of data-driven photometric stereo methods. Virtual Reality & Intelligent Hardware, 2, 213–221.