Multimodal Light-Field Camera with External Optical Filters Based on Unsupervised Learning

Takumi Shibata, Fumihiko Sakaue and Jun Sato
Nagoya Institute of Technology, Nagoya, Japan

Keywords: Light Field Camera, Multi-Modal Imaging, External Optical Filters.

Abstract: In this paper, we propose a method of capturing multimodal images in a single shot by attaching various optical filters to the front of a light-field (LF) camera. However, when a filter is attached to the front of the lens, the result of capturing images from each viewpoint will be a mixture of multiple modalities. Therefore, the proposed method uses a neural network that does not require prior learning to analyze such a modal mixture image to generate an image of all the modalities at all viewpoints. By using external filters as in the proposed method, it is possible to easily switch filters and realize a flexible configuration of the shooting system according to the purpose.

1 INTRODUCTION

In recent years, images that record various information about a target object, such as multispectral images, polarization images, and high dynamic range (HDR) images, have been in the spotlight. Since these images contain information that is difficult to handle in ordinary RGB images, they can be applied to various applications by using them for different purposes. If these information could be acquired and used simultaneously, it would be possible to construct an image processing system with higher accuracy. However, these images often require a dedicated camera, and it is difficult to simultaneously acquire this information from the same viewpoint. Therefore, in this paper, we investigate a method of acquiring multimodal images with a single shot from a single camera. Several methods for acquiring multimodal images using a single camera have been studied in the previous years (Horstmeyer et al., 2009), often using image sensors equipped with special filter arrays. This makes it possible to acquire various information depending on the filter, such as multispectral image (Xie et al., 2019), HDR image, and polarization information image. However, the filter configuration cannot be changed immediately with such a special image sensor, and flexible operation such as changing the modality according to the purpose is not possible. Therefore, we propose a method for acquiring multimodal images using external filters that can be replaced or combined with other filters.

In order to acquire various information simultaneously from a single viewpoint, we focus on a light field (LF) camera, which is a camera that acquires 4D LF images. The camera can obtain images (subaperture images) equivalent to those taken by multiple cameras at the same time by taking a single shot. In this study, we consider the simultaneous acquisition of images through different filters in a single shot by attaching various filters to each area of the main lens of the LF camera and taking a picture, as shown in Fig.1. However, when the filters are mounted on the front of the camera, the filter position is different from the optical center, so the ideal group of images as shown in Fig.1 cannot be obtained directly. Several modalities are mixed in each subaperture image, and mixed modal information is lacking. In this research, we aim to analyze such mixed-modal images taken from multiple viewpoints to capture multimodal images with no missing parts in a single shot by a single camera.

2 MIXED-MODAL IMAGING USING LFcamera

First, we describe the light field (LF) camera used in this study, which directly records four-dimensional information on light rays in a target scene. In this study, we use a plenoptic camera (Ng et al., 2005) that realizes LF capturing with a single camera by at-

Shibata, T., Sakaue, F. and Sato, J.
Multimodal Light-Field Camera with External Optical Filters Based on Unsupervised Learning.
DOI: 10.5220/0011801900003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 4: VISAPP*, pages 483-490
ISBN: 978-989-758-634-7; ISSN: 2184-4321

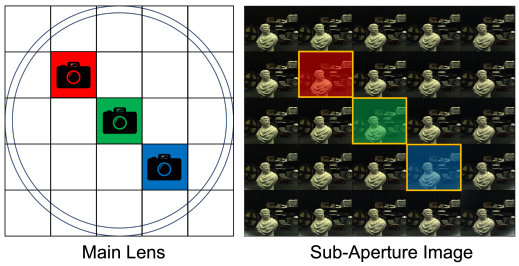


Figure 1: External multi-modal filters for light-field camera and ideal sub-aperture images.

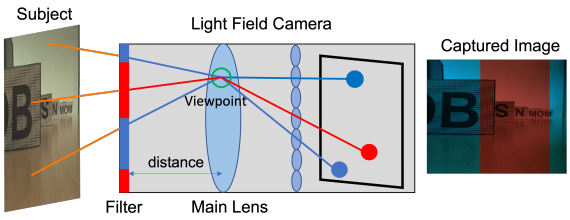
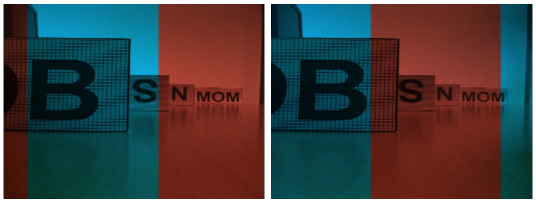


Figure 2: Image capturing with external multi-modal filters.

taching a microlens array to the image sensor. We consider what kind of images can be obtained when an external filter that allows light rays with different characteristics to pass through each part of the camera is attached to this plenoptic LF camera, as shown in Fig.2. Now, let us focus on a certain region of the main lens (subaperture) as shown in Fig.2, and consider the subaperture image obtained by light rays passing through this region. In this case, as shown in the figure, the lens and filter plane are located at different positions, so each light ray passes through a different filter depending on the direction of incidence and reaches the main lens. Therefore, even if a subaperture image is generated from the acquired LF, it is not possible to obtain a single modal image corresponding to the filter in front of the subaperture, and thus, various modalities are mixed in the image. In this study, we call the image a mixed-modal image. Since the rays passing through each subaperture change according to the position of the subaperture, the modal mixed image changes depending on the position of the subaperture, as shown in Fig.3. The proposed method estimates all unmixed modal images at all subapertures from the set of such modal mixed images.



(a) Left Viewpoint Image (b) Right Viewpoint Image

Figure 3: Difference of mixed modal image according to the viewpoint.

3 MULTI-MODAL IMAGE ESTIMATION BY DEEP IMAGE PRIOR

3.1 Deep Image Prior

In this study, we utilize Deep Image Prior(Ulyanov et al., 2017) to estimate multi-modal images. Deep Image Prior uses neural networks as prior knowledge (Prior) in image generation to achieve natural image generation. In Deep Image Prior, a noise image N is input to a neural network with a U-Net structure, and the output image x is obtained as follows:

$$x = P(\theta, N) \tag{1}$$

where, $P(\theta, N)$ is output from U-net when the noise N is input to the network and θ is a set of parameters in the network. Deep Image Prior can generate various images by changing the parameter θ of the CNN. For example, to make the output x closer to the target image x_o, the parameter θ is optimized by as follows:

$$\theta^* = \arg\min_{\theta} ||(P(\theta, N) - x_o)||^2 \tag{2}$$

The reconstructed image $P(\theta^*, N)$ can be obtained from the parameter θ^* obtained in this way. By changing this evaluation function according to the purpose, various image processing such as image super-resolution and inpainting can be realized without prior learning(Ho et al., 2021)(Rasti et al., 2022).

3.2 Conditions for Image Generation

Next, we consider the constraints imposed on Deep Image Prior to generate multi-modal images. In this study, we focus on the fact that the modality information that can be obtained from each pixel changes for each viewpoint in generating each modal image. Figure3 shows an example of a subaperture image obtained by using a red and blue mixture filter, and it can be seen that rays of light passing through different filters are obtained for each viewpoint, even when the same object is captured. This is because the relative positional relationship with the filter changes for each subaperture. Therefore, if the image from each viewpoint can be transformed into an image from a different viewpoint, it is possible to acquire information on various modalities from all viewpoints. Therefore, a viewpoint transformation network is constructed to represent the change in image with viewpoint transformation and used for image generation.

Assuming that there is a certain correlation between the modal images, it is possible to transform one modal image to another using a neural network.

Therefore, a modal transformation neural network that transforms each modal image into another modal image is also trained, and this is also used as a condition for image generation. Furthermore, by sharing some of these networks, a multi-task learning framework is applied to estimate multimodal images with high accuracy.

In the following, for the simplification of the discussion, we will consider the case where images with different modals are captured using filters that transmit light of red wavelength and blue wavelength, respectively, as shown in Fig.3. Let I_i be the modal mixed image taken at viewpoint i. It is also assumed that at each viewpoint, it is known which modal information was acquired at which pixel, and that a mask M_i^R replacing the non-red region with 0 and a mask M_i^B replacing the non-blue region with 0 are obtained. Under these conditions, the objective is to bring the output $\hat{I}_i^R = P(\theta_i^R, N)$ and $\hat{I}_i^B = P(\theta_i^B, N)$ obtained by Deep Image Prior to their respective modal images. In this section, we consider, in particular, the constraints available when focusing on $\hat{I}_i^R$.

3.3 Image Inpainting Constraint

First, consider the constraint that image inpainting is used to generate an image from an image of the target modal contained in a modal mixture. In this case, the error in the unmasked area becomes the evaluation function for image generation, as follows:

$$\varepsilon_P = ||M_i^R \hat{I}_i^R - M_i^R I_i||^2 \quad (3)$$

Minimizing this evaluation function yields an image that mimics the input for regions where red information can be captured directly, and an interpolated image based on the input for regions where it cannot be obtained.

3.4 Constraint from Viewpoint Transformation

Next, we consider using information from images taken from different viewpoints by estimating the disparity between the images. In this study, we extend the method of Luo et al(Luo et al., 2018). to estimate the disparity. In this method, multiple images are prepared for the image to be viewpoint transformed that have been shifted by k pixels in advance, and these are called the shifted image set $S^k(I)$. A weight map W^k representing the weight of each pixel is estimated for each image in this shifted image group, and the weighted average of the weight maps is computed to generate the viewpoint transformation image. The weight map W^k indicates from which shifted image

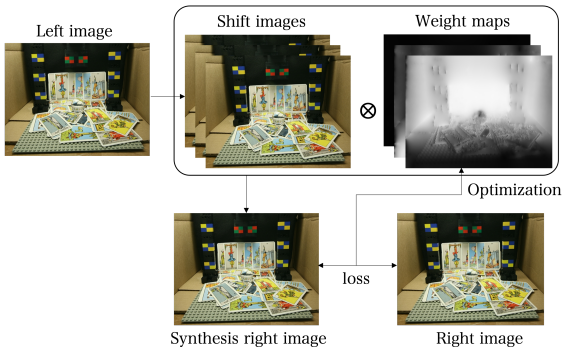


Figure 4: Overview of the Viewpoint Transformation Using Shifted Image Set.

the pixel values are referenced for each pixel in the viewpoint-transformed image, and by optimizing this map, an image that is appropriately shifted according to disparity can be generated.

In the method of Luo et al. as shown in Fig.4, the viewpoint transformation is performed according to the input by learning the relationship between the input and W^k in advance. In this study, the viewpoint transformation is performed by generating this weight map using Deep Image Prior. However, since the images handled in this study are modal mixed images, the number of pixels that can be compared is limited when directly comparing input images. Therefore, the viewpoint-transformed image is estimated by comparing the generated image $\hat{I}_i^m (m \in \{R, B\})$ at viewpoint i with the generated image $\hat{I}_j^m$ at viewpoint j. This optimizes the weight map W_j^k, which represents the k pixel shift, for the conversion of the viewpoint j image to the viewpoint i image as follows:

$$\varepsilon_{V_{j\to i}} = \sum_{m\in\mathcal{M}} ||\hat{I}_i^m - \sum_k W_j^k S^k(\hat{I}_j^m)||^2 \quad (4)$$

where $\mathcal{M} = \{R, B\}$. When the generated image is fixed, this function is an evaluation function for the viewpoint transformation. On the other hand, if W is fixed and the generated image is variable, it becomes a constraint on image generation considering the result of viewpoint transformation. In this study, the optimization of W and $\hat{I}$ is performed simultaneously to simultaneously perform viewpoint transformation and image generation.

3.5 Modality Transformation by Neural Network

Next, consider how to use images from other viewpoints and other modalities in image generation. As shown in Fig.3, since each mixed-modal image is taken from a different viewpoint (subaperture), the filter pass points are different. Therefore, there are over-

lapping regions between the modal regions. By learning the correspondence between each modal from these overlapped regions, we can transform from one modal to another.

Now, consider a modality transformation T_R from blue to red using the blue region $M_i^B I_i$ from the i-th viewpoint and the red region image $M_j^R I_j$ from the j-th viewpoint. In this case, the loss function $\varepsilon_{B\to R}$ used for learning is defined as follows:

$$\varepsilon_{B\to R} = \sum_{j(\neq i)} ||M_i^B T_R(M_i^B I_i) - M_i^B V_{j\to i}(M_j^R I_j)||^2 \quad (5)$$

where $V_{j\to i}$ is the function that transforms the j viewpoint image to the i viewpoint image by the viewpoint transformation described earlier. The transformation function T_R obtained by optimizing this loss function is used to transform a blue modal region into a red modal region. This yields an image $T_R(M_i^B I_i)$ that predicts the red image from the blue image. Using this, the following evaluation function is added to the Deep Image Prior output $\hat{I}_i^R$.

$$\varepsilon_T = ||M_i^B \hat{I}_i^R - M_i^B T_R(M_i^B I_i)||^2 \quad (6)$$

By adding this loss function and optimizing $\hat{I}_i^R$, information from another modal image can be used for image generation.

3.6 Multi-Modal Image Estimation

Simultaneous optimization of the above evaluation functions produces the target single-modal image. The aforementioned evaluation functions include various functions such as viewpoint transformation, modal transformation, etc., all of which use the same image as input. Therefore, by minimizing the sum of all of these evaluation functions, it is possible to generate an image that satisfies all conditions. Therefore, the evaluation function ε_i^R for estimating $\hat{I}_i^R$ is expressed as follows:

$$\varepsilon_i^R = \varepsilon_P + \sum_{j(\neq i)} \varepsilon_{V_j} + \varepsilon_{B\to R} + \varepsilon_T \quad (7)$$

In addition, this evaluation function includes the generated images other than the R-modal images from the i viewpoints. Therefore, in actual optimization, all viewpoints and all modal images are estimated simultaneously by minimizing ε shown as follows:

$$\varepsilon = \sum_i \sum_{m\in\mathcal{M}} \varepsilon_i^m \quad (8)$$

At last, we can estimation of multimodal images for all viewpoints from any combination of multimodal filters using only the information from the input images.

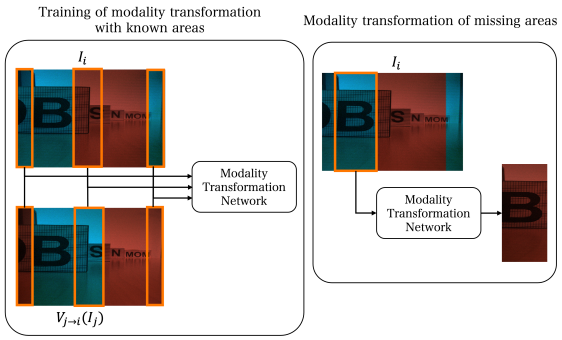


Figure 5: Modality transformation using neural networks.

4 RESULTS

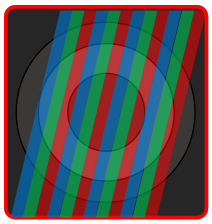
4.1 Using Color Filters

Semi-simulation experiments were conducted to obtain multimodal images using the proposed method. In the proposed method, by combining as many fine-grained filters as possible, the overlapped area of each filter increases, and the estimation accuracy of the multimodal image can be improved. However, since it is difficult to create such fine-grained multimodal filters, we first confirmed the effectiveness of the proposed method in a semi-simulation experiment. In this experiment, filters covering the entire surface of the lens were mounted, and images were acquired through each filter. The captured images were then combined according to the assumed filter shapes to obtain a mixed-modal image. in Fig.6(a) the shape of the filters is shown, with the red, blue, and green color filters arranged on a slant in front of the lens. The subaperture images taken from the multimodal filters arranged in this way are estimated as shown in Fig.8. We evaluated the proposed method by comparing the results obtained by the proposed method with those of each of the modalities used to generate the input images.

The estimated disparity image resulting from the generation of the center viewpoint multimodal image by the proposed method and the ground truth image obtained are shown in Fig.13. The results confirm the effectiveness of the proposed method, as the proposed method estimates an image that is very similar to the ground truth.

On the other hand, the details of the image are not sufficiently restored, resulting in a slightly blurred image overall. This may be due to the fact that the deep image prior requires a large number of updates to represent high-frequency portions of the image, which can be improved by adjusting the network structure and the number of optimization cycles. The RMSE

per pixel of the true value image and the estimated image is kept low at 5.4 on average, as shown in Tab.1, confirming that the image is generated appropriately.

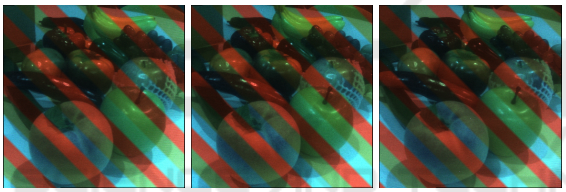


(a) Simulation

(b) Real Environment

Figure 6: Light field camera and simulated multimodal filters.

Figure 7: Target scene.



(a) Left image (b) Center image (c) Right image

Figure 8: Input image synthesized from ground truth images.

Table 1: Evaluation of the proposed method by RMSE.

Modality of the image	RMSE
Red	5.122
Green	5.809
Blue	5.385

4.2 HDR Image Synthesis

Next, we show the results of generating HDR images using the proposed method. In this experiment, images were captured assuming that three different types of light-reducing filters with different degrees of light-reduction were installed. As in the previous experiment, each filter was placed on a tilt. Each image was captured at three different brightness levels by adjusting the exposure time.

The input image synthesized from these images is shown in Fig.9. Based on this image, we estimated different brightness images using the proposed method. The brightness of each image is estimated by the proposed method. The HDR image is obtained by merging these images. The estimated image and the true value image are shown in Fig.14. The results show that the HDR image generation was successfully performed from a single-shot image, as the areas where white skipping and blacking out occurred were well represented in each image.

(a) Left image (b) Center image (c) Right image

Figure 9: Input images synthesized from ground truth images.

(a) Left image (b) Center image (c) Right image

Figure 10: Input images synthesized from ground truth images.

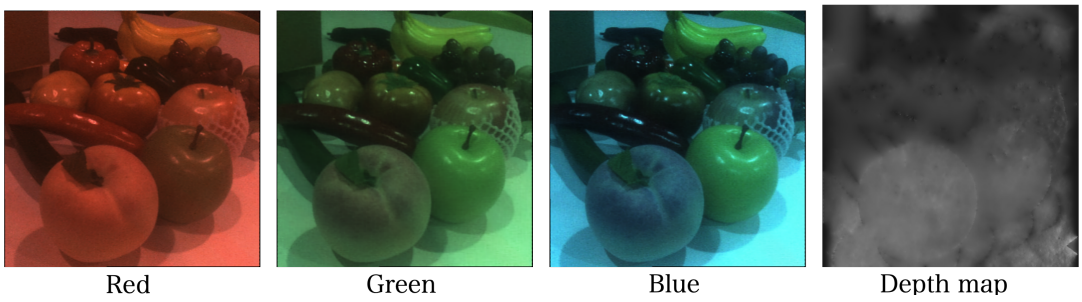
4.3 Polarization Image Estimation

Next, we show the results of generating polarization images using the proposed method. In this experiment, a multimodal filter with a diagonal combination of polarization filters in four directions (0°, 45°, 90°, and 135°) was attached to the front of the LF camera, and a modal mixture of images obtained by simulation was combined. We use this image as an input image to estimate the four-directional polariza-

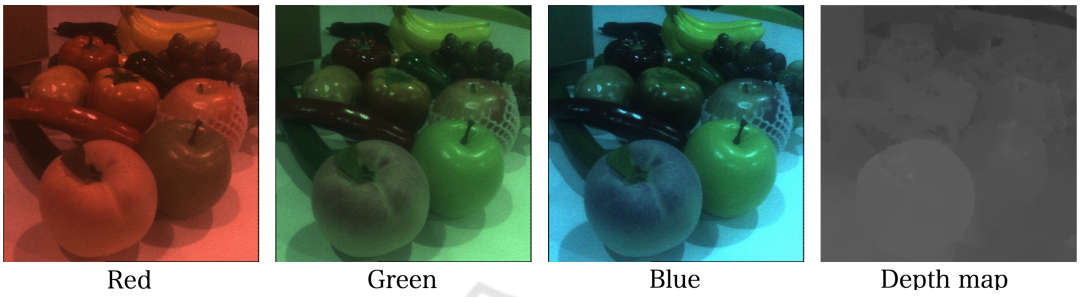
Figure 11: Target scene.

(a) Left image (b) Center image (c) Right image

Figure 12: Input images.



(a) Synthesized images



(b) Ground truth images

Figure 13: Synthesized images and ground truth.

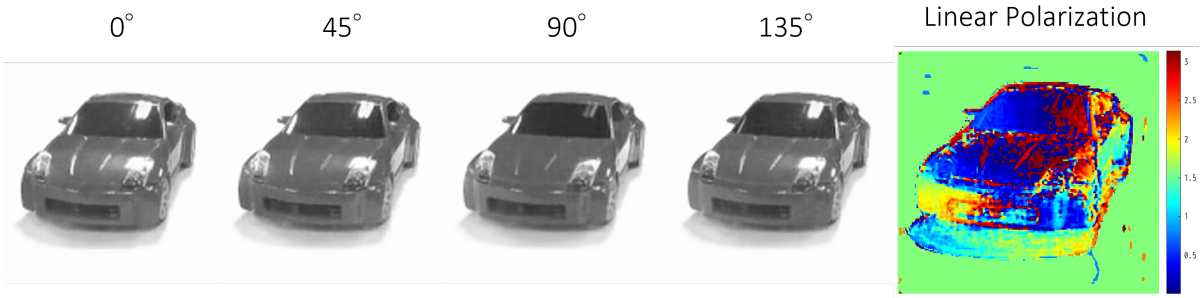
(a) Synthesized images

(b) Ground truth images

Figure 14: Synthesized images and ground truth.

tion image. Furthermore, the polarization angle maps were estimated using the polarization images in the four directions and compared with the true values to evaluate each of the polarization images generated.

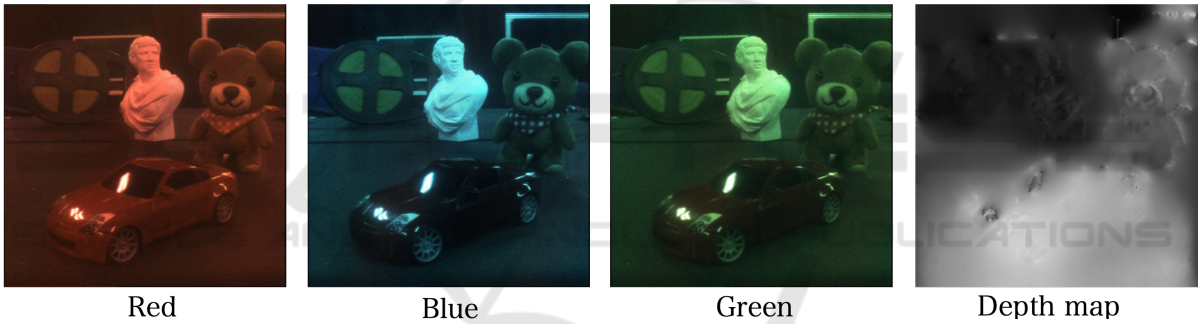
The input images synthesized by the simulation are shown in Fig.10, and the results of the estimation using the proposed method and the true value images using these images are shown in Fig.15. The results show that the proposed method can generate images that are close to the true value for each angle. In particular, focusing on the polarization image of 90°, it can be confirmed that, unlike the other polarization images, the reflection of light is reduced and an image close to the generated image and the true value image can be generated. Comparing the polarization angle maps, it can be confirmed that although the estimation accuracy in the high-frequency region is lower, the overall image is similar to the true value, indicating that appropriate image estimation has been achieved.



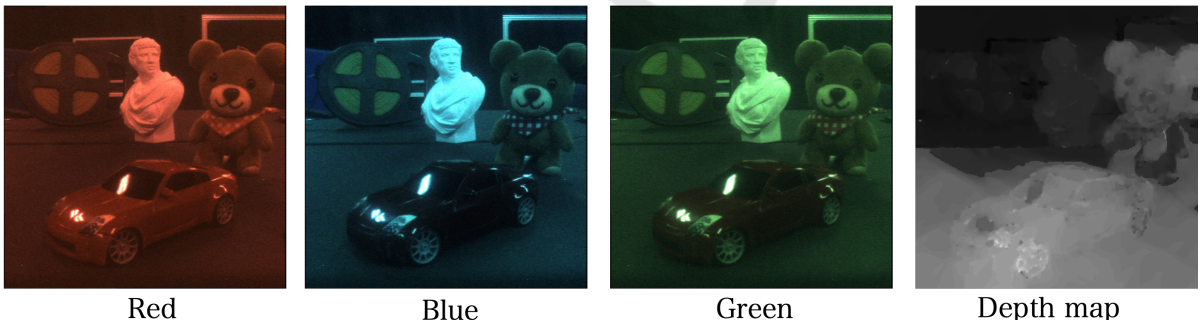
(a) Synthesized images

(b) Ground truth images

Figure 15: Synthesized images and ground truth.



(a) Synthesized images



(b) Ground truth images

Figure 16: Synthesized images and ground truth.

4.4 Real-World Experiments Using Color Filters

Finally, we show the results of a real-world experiment in which the color multimodal filters that we actually created were attached. In this experiment, a multimodal filter consisting of red, blue, and green color filters cut into strips and arranged diagonally on the front of the LF camera is attached to the camera as shown in Fig.6(b). A portion of the modal mixture of images obtained by this method is shown in Fig.12. In this experiment, a group of images from 25 verti-

cal and horizontal viewpoints, including the images in Fig.12, were used as input images to estimate the images obtained by passing through each filter.

The multimodal and disparity images of the central viewpoint generated by the proposed method and their true-value images are shown in Fig.16. The results show that the proposed method is effective even with the actual multimodal filter, since it can be confirmed that the images generated are close to the true-value images.

The RMSE values averaged 9.6, confirming that the estimation accuracy is lower than in the color filter simulation experiment. However, the estimation accuracy in the real environment experiment is expected to improve by attaching a multimodal filter with a more ideal shape.

Table 2: Evaluation of the proposed method by RMSE.

Modality of the image	RMSE
Red	9.448
Green	10.015
Brue	9.752

5 CONCLUSION

In this paper, we propose a method for acquiring multimodal images based on unsupervised learning using an LF camera and exterior filters. This method is expected to have a wide range of applications because it is easy to switch acquisition modalities according to the purpose and does not require any training data.

REFERENCES

Ho, K., Gilbert, A., Jin, H., and Collomosse, J. (2021). Neural architecture search for deep image prior. *Computers & Graphics*, 98:188–196.

Horstmeyer, R., Euliss, G., Athale, R., and Levoy, M. (2009). Flexible multimodal camera using a light field architecture. *2009 IEEE International Conference on Computational Photography, ICCP 09*, pages 1 – 8.

Luo, Y., Ren, J. S. J., Lin, M., Pang, J., Sun, W., Li, H., and Lin, L. (2018). Single view stereo matching. *CoRR*, abs/1803.02612.

Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., and Hanrahan, P. (2005). Light Field Photography with a Hand-held Plenoptic Camera. Research Report CSTR 2005-02, Stanford university.

Rasti, B., Koirala, B., Scheunders, P., and Ghamisi, P. (2022). Undip: Hyperspectral unmixing using deep image prior. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. S. (2017). Deep image prior. *CoRR*, abs/1711.10925.

Xie, Q., Zhou, M., Zhao, Q., Meng, D., Zuo, W., and Xu, Z. (2019). Multispectral and hyperspectral image fusion by MS/HS fusion net. *CoRR*, abs/1901.03281.