IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 December 2022, accepted 27 January 2023, date of publication 1 February 2023, date of current version 8 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3241348

RESEARCH ARTICLE

Monocular Depth Estimation of Old Photos via Collaboration of Monocular and Stereo Networks

JU HO KIM[1], KWANG-LIM KO[2], LE THANH HA[3], (Member, IEEE), AND SEUNG-WON JUNG[1], (Senior Member, IEEE)

[1]Department of Electrical Engineering, Korea University, Seoul 02841, South Korea
[2]Department of Automotive Convergence, Korea University, Seoul 02841, South Korea
[3]University of Engineering and Technology, Hanoi 100000, Vietnam

Corresponding author: Seung-Won Jung (swjung83@korea.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT Future Planning, under Grant NRF-2020K1A3A1A05103460. The work of Le Thanh Ha was supported by the Korea-Vietnam Joint Research Project through the AI-Based Application for 3D-Model Restoration of Vietnamese and Korean Cultural Heritage under Grant NĐT/KR/21/17.

ABSTRACT Old photos that were captured about a century ago have archaeological and historical significance. Many of the old photos have been successfully digitized, but most of them suffer from severe and complicated distortion. Thus, prior studies have focused on image restoration tasks such as denoising, inpainting, and colorization. In this paper, we pay attention to the depth estimation of old photos, enabling a more enjoyable appreciation of them and helping better understand past human life, activities, and environments. Because most old photos are available as single-view images, monocular depth estimation techniques can be considered a solution. However, most high-performance techniques are based on supervised learning, which requires ground-truth depth maps. Because this kind of supervised learning is not feasible for old photos, in this paper, we present a learning framework that finetunes a pretrained monocular depth estimation network for each old photo. Specifically, the pretrained monocular depth estimation network predicts stereo depth maps for stereo image rendering. Then, the pretrained stereo network predicts depth estimates from the rendered stereo image pair. By extracting reliable depth estimates and using them for supervision of the monocular network, the monocular network can be gradually learned to produce a high-quality depth map of the given old photo. From the qualitative and quantitative performance evaluations on old photos, we demonstrate the effectiveness of the proposed method.

INDEX TERMS Knowledge distillation, monocular depth estimation, old photo, zero-shot learning.

I. INTRODUCTION

Since the invention of the camera, photographs have been used as a means of visual communication and expression. Nowadays, many people are sharing their photos with others through social media. However, this did not apply a hundred years ago when cheap and comfortable cameras were not available to consumers. Fortunately, there are still many valuable photos taken with low-performance cameras in such a period, which play an important role in understanding past human life, activities, and environments. These photos are monochromatic and have low signal-to-noise ratios (SNRs). In addition, since they have been digitized recently, they suffer from multiple and complicated distortions due to aging and physical damage. Hence, understanding 3D geometric features, *i.e.*, depth estimation, is challenging in these degraded old photos.

Photographic restoration for old photos can alleviate the difficulty of understanding the 3D geometric features in challenging old photos. Although there is mixed degradation in old photos, including scratches and blotches, film noise, and the lack of color, most prior studies address each degradation separately. In particular, the inpainting task received the most interest, which requires two steps: identification of scratches and blotches and recovery of these damaged areas using the textures from the vicinity [3], [4] or external images [5], [6]. One recent work [2] addresses the mixed degradation by introducing latent space mapping with synthetic paired data,

The associate editor coordinating the review of this manuscript and approving it for publication was Guillermo Botella Juan.

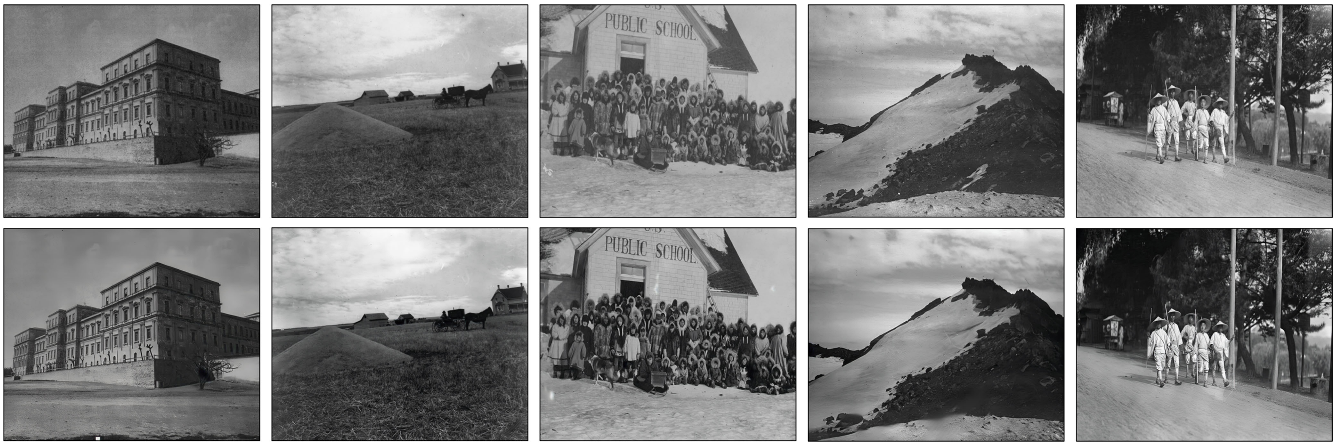


FIGURE 1. Several examples of the old photos (top) [1] and their restoration results (bottom) obtained using [2].

which enables an automatic repair of old photos. Although photographic restoration of old photos has been extensively studied, geometric reconstruction of old photos have received less attention.

In this paper, we pay attention to the estimation of geometric information from old photos. In particular, we attempt to estimate a depth map corresponding to each old photo to enable a more realistic rendering of the previous moment. Since video frames or images from different viewpoints are hardly available for old photos, estimation of the depth map from a single photo, referred to as monocular depth estimation, is required to achieve our objective. Monocular depth estimation has been extensively studied in the last decades, and several state-of-the-art techniques [7], [8], [9], [10], [11] show very promising results on specific scenes, *e.g.*, road scenes. However, these techniques are based on supervised learning, which requires ground-truth depth maps for training. Because such ground-truth depth maps are not available for old photos, an alternative approach is required.

Our approach is to train a monocular depth estimation network with the help of a stereo depth estimation network. Given the monocular depth estimation network pretrained on modern photos, we first obtain a pair of left-view and right-view depth maps of an old photo and then use them to render a stereo image pair. The rendered stereo image and the stereo depth estimation network are used to obtain another pair of left-view and right-view depth maps. We then extract only reliable depth estimates by the left-right consistency check and finetune the monocular depth estimation network on each old photo. Our contributions are three-fold:

- We present a novel learning framework that enables finetuning of the monocular depth estimation network without requiring any ground-truth supervision signals.
- We demonstrate the superiority of our method over the state-of-the-art unsupervised monocular depth estimation methods on old photos.
- We introduce a pairwise comparison-based quality evaluation method for the depth images estimated from old photos.

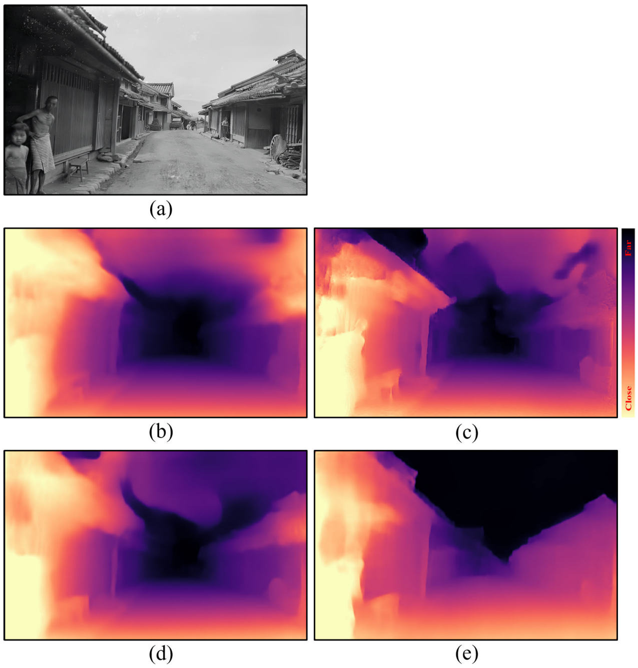


FIGURE 2. Example of monocular depth estimation results: (a) Input image, (b) depth map obtained using CADepth-Net [9], (c) depth map obtained using HRDepth [10], (d) depth map obtained using MonoDepth2 [11], and (e) depth map obtained by the proposed method.

The rest of this paper is organized as follows: Section II reviews previous research related to our work; Section III describes the proposed method; Section IV presents the implementation details and experiment results; Section V concludes the paper.

II. RELATED WORKS

A. OLD PHOTO RESTORATION

Old photos that were captured about a century ago suffer from multiple and complicated distortions, which naturally require multiple steps of image restoration to enhance their visibility. However, most previous studies on old photos focus on image

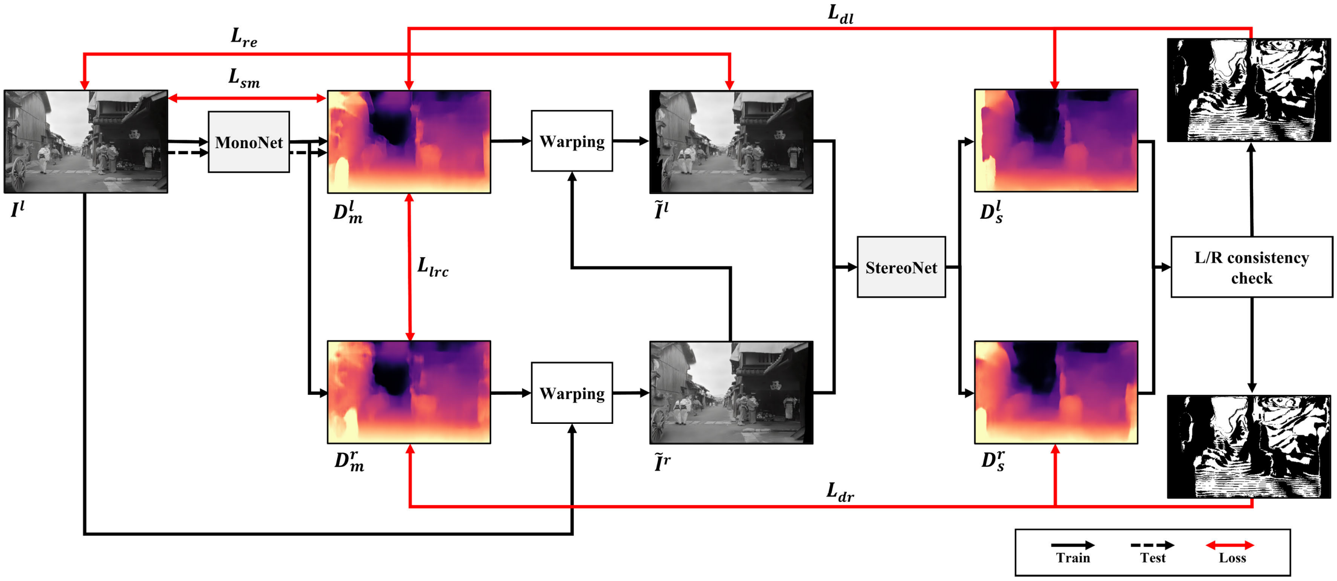


FIGURE 3. Overview of the proposed framework. From a single old photo I^l, the pretrained MonoNet predicts left-view and right-view depth maps D^l_m and D^r_m. A warped right-view image $\tilde{I}^r$ is synthesized from I^l and D^r_m. A warped left-view image $\tilde{I}^l$ is synthesized from the $\tilde{I}^r$ and D^l_m. Given stereo images $\tilde{I}^l$ and $\tilde{I}^r$ as input, StereoNet predicts left-view and right-view depth maps D^l_s and D^r_s. After checking the left-right consistency on both depth maps D^l_s and D^r_s, the consistent depth pixels from StereoNet are used to supervise MonoNet.

inpainting [3], [4], [12] because damaged pixels, such as scratches and blotches, need special treatment. One exceptional recent study [2] restores multiple distortions simultaneously by using an image translation framework. Specifically, synthetically distorted image pairs are used to help map old photos to the latent space, and a global branch with a non-local block [13] is used with residual blocks for enabling multiple degradation restoration during latent space transformation. Fig. 1 shows several examples of the old photos used in our study [1] and their restoration results [2]. The quality of old photos can be significantly improved by [2], and thus we use these pre-processed old photos for our monocular depth estimation task.

B. STEREO DEPTH ESTIMATION

Classical stereo matching algorithms are extensively studied to find correspondence between stereo images by calculating matching costs [14], [15], [16], [17]. With the development of learning-based methods, convolutional neural network (CNN)-based stereo depth estimation methods demonstrate superior performance on challenging real-world scenes [18], [19], [20], [21], [22]. Based on characteristics that both stereo matching and optical flow aim at finding pixel correspondence, some approaches tried incorporating optical flow information to estimate more robust stereo depth maps [23], [24], [25]. Due to the availability of binocular depth cue, stereo depth estimation is generally more robust and accurate than monocular depth estimation at the expense of increased computational costs and necessity of stereo images. Notably, for scenes of old age, stereo images are not available. Nevertheless, we generate stereo images of old scenes using a warping module to exploit them as source images for pseudo-ground-truth depth maps.

C. MONOCULAR DEPTH ESTIMATION

Traditional monocular depth estimation techniques rely on monocular depth cues such as texture, shade, focus, and occlusion for estimating a depth map from a single image. However, monocular depth cue-based methods suffer from inherent ambiguities when applied to unconstrained scenes. Due to the remarkable progress of deep learning, learning-based monocular depth estimation methods are currently dominating [26]. If the corresponding pairs of images and depth maps are available, a CNN can be trained to predict the depth map from the icuest image. For example, the pioneering work of Eigen et al. [27] presented a coarse-to-fine framework that obtains a coarse global prediction and a fine local prediction sequentially. Many of the follow-up studies attempted to improve the performance by changing network architectures and loss functions [28], [29], [30]. For the input old photo shown in Fig. 2(a), Figs. 2(b), (c) and (d) show the depth maps obtained by the state-of-the-art methods [9], [10], [11] trained on the KITTI dataset. Due to the unavoidably large domain gap between the training and test images, depth maps are obtained with large errors. When ground-truth depth maps do not exist for the target domain as in our old photo application, depth estimation networks need to be trained in an unsupervised manner. If the left and right images are given, a network can first output the depth map for the left image. A new left image can then be reconstructed by warping the right image using the obtained depth map and compared with the original left image for training the network [31]. A similar approach can be applied to monocular video sequences [32], [33]. However, these scenarios are not applicable to old photos available as single images.

D. KNOWLEDGE DISTILLATION

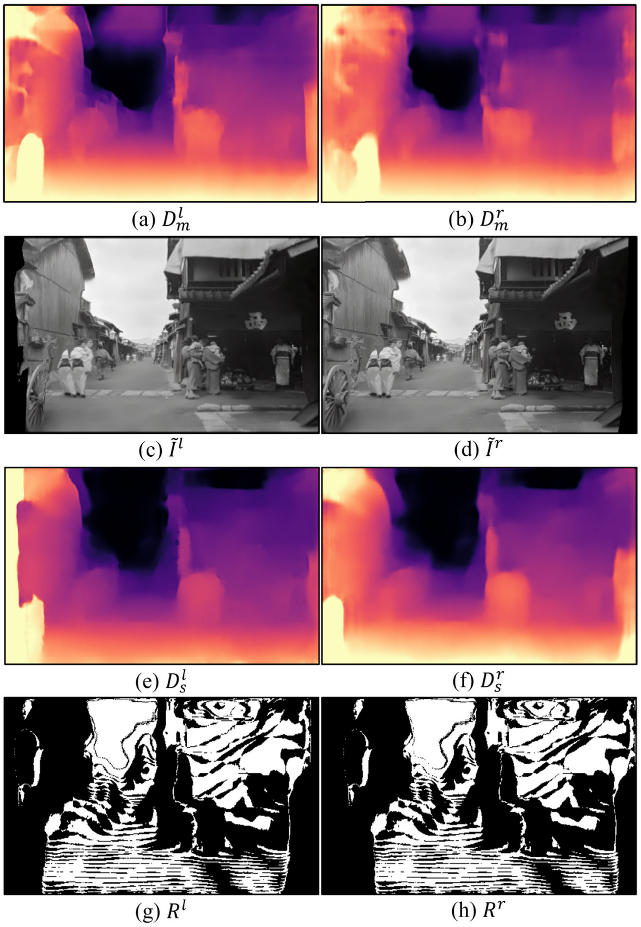
Knowledge distillation, which has advantages in increasing performance and reducing model complexity, has been actively studied recently. The student-teacher structure, proposed by Hinton et al. [34], is based on a strategy that a more accurate and deeper network can serve as a teacher model to guide the less complex student model to estimate better results. In the field of depth estimation, stereo networks that take left-view and right-view images as input usually produce more accurate depth maps than monocular networks. Accordingly, stereo and monocular networks have been employed as teacher and student networks, respectively, for the monocular depth estimation task. Some studies employed knowledge distillation by making monocular networks follow the predictions of stereo networks [35], [36], [37]. In addition to the depth maps, the intermediate feature maps extracted from the stereo networks have also been used as supervision signals for monocular networks [38], [39], [40]. Several other studies attempted to use only reliable predictions of the stereo network for knowledge distillation [41], [42]. In line with these recent works, we present the student-teacher strategy dedicated to the depth estimation of old photos.

III. PROPOSED METHOD

Due to the lack of ground-truth depth maps or multi-view views, depth maps for old photos cannot be accurately estimated by existing monocular depth estimation methods. Therefore, inspired by the knowledge distillation, we propose to train the monocular network with the pseudo-ground-truth depth maps generated by the stereo network. The generated pseudo-ground-truth depth maps are refined through the left-right consistency check and used to finetune the monocular network. Fig. 3 shows an overall framework of the proposed method, where MonoNet and StereoNet represent the monocular and the stereo depth estimation networks, respectively. Throughout the paper, 'depth' specifically implies 'disparity.' All the old photos are used after preprocessed by the old photo restoration technique [2].

A. PRETRAINING OF MONOCULAR AND STEREO NETWORKS

Our framework is not restricted to specific architectures of StereoNet and MonoNet, and we used DispNetC [43] and VGG-16 [44] for our experiment. The overall pretraining process of StereoNet and MonoNet is the same as [35], but the output layers are modified to obtain both left-view and right-view depth maps. Specifically, StereoNet is supervised by the modern photo dataset [43] where ground-truth depth maps are available. After the training of StereoNet, MonoNet is trained in an unsupervised manner by using the predictions of StereoNet as pseudo-ground-truth. The training datasets and implementation details will be explained in Sections IV-A and IV-B, respectively.


(a) D_m^l (b) D_m^r

(c) $\tilde{I}^l$ (d) $\tilde{I}^r$

(e) D_s^l (f) D_s^r

(g) R^l (h) R^r

FIGURE 4. Example of monocular depth estimation results: (a) and (b) are the estimated depth maps from MonoNet, (c) and (d) are the warped left-view and right-view images, (e) and (f) are the estimated depth maps from StereoNet, (g) and (h) are the reliability maps obtained by the left-right consistency check.

B. OVERALL FRAMEWORK

Let I^l denote a given old photo corresponding to the left-view. Likewise with other monocular depth estimation networks [45], [46], MonoNet produces both left-view and right-view depth maps, denoted as D_m^l and D_m^r, respectively. Figs. 4(a) and (b) show examples of the depth maps obtained by the pretrained MonoNet for the input image shown in Fig. 2(a). Due to a large domain gap between the dataset used for training [47] and the input old photo, the depth maps are obtained with significant errors. Given I^l and D_m^r, we can synthesize the right-view old photo $\tilde{I}^r$ as follows:

$$\tilde{I}^r = W\left(I^l, D_m^r\right), \tag{1}$$

where W represents the backward warping function, and we used the differentiable warping layer [45] in our implementation. Similarly, given $\tilde{I}^r$ and D_m^l, we can synthesize the left-view old photo $\tilde{I}^l$ as follows:

$$\tilde{I}^l = W\left(\tilde{I}^r, D^l_m\right). \tag{2}$$

Figs. 4(c) and (d) show examples of the synthesized stereo images, $\tilde{I}^l$ and $\tilde{I}^r$. Because of a large domain gap between modern and old photos, the depth maps are obtained with significant errors, and consequently, the stereo images contain artifacts. The parameter updates of MonoNet are essential for handling old photos, but the lack of ground-truth depth maps makes MonoNet training challenging.

Our solution is to exploit StereoNet, which is proven to be less sensitive to the domain gap between training and test datasets [35]. From $\tilde{I}^l$ and $\tilde{I}^r$, StereoNet can produce left-view and right-view depth maps, denoted as D^l_s and D^r_s, respectively. By checking the left-right consistency [45], we can easily obtain reliability maps of the left-view and right-view, denoted as R^l and R^r, respectively. The threshold for the consistency check is set to 0-pixel distance. Figs. 4(e) and (f) show the depth maps obtained by applying $\tilde{I}^l$ and $\tilde{I}^r$ to the pretrained StereoNet. Because of low-quality synthesized stereo images, the depth maps are obtained with large errors. However, we can still obtain some reliable estimates that are valuable for finetuning MonoNet. Specifically, we use the pixels with value ones in the reliability maps, as shown in Figs. 4(g) and (h), for the supervision of MonoNet finetuning. Let Z_{Φ_m} and Z_{Φ_s} denote MonoNet and StereoNet parameterized by Φ_m and Φ_s, respectively. The two pairs of stereo depth maps are obtained as follows,

$$\left\{D^l_m, D^r_m\right\} = Z_{\Phi_m}\left(I^l\right), \tag{3}$$

$$\left\{D^l_s, D^r_s\right\} = Z_{\Phi_s}\left(\tilde{I}^l, \tilde{I}^r\right). \tag{4}$$

For each old photo I^l, we finetune MonoNet for T iterations and obtain the depth map from the finetuned MonoNet.

C. LOSS FUNCTIONS

The training loss function L_{total} is defined as

$$L_{total} = (L_{dl} + L_{dr}) + \mu_1 L_{re} + \mu_2 L_{sm} + \mu_3 L_{lrc}, \tag{5}$$

which is a combination of four main terms: the distillation loss, the reconstruction loss, the depth smoothness loss, and the left-right consistency loss. The weighting factors were empirically chosen as $\mu_1 = 0.5$, $\mu_2 = 0.1$, and $\mu_3 = 0.01$.

The distillation loss from StereoNet to MonoNet for the left-view, denoted as L_{dl}, is defined as follows:

$$L_{dl}\left(D^l_m, D^l_s\right) = \frac{1}{M}\left\|D^l_m(P) - D^l_s(P)\right\|, \tag{6}$$

where $\|\cdot\|$ measures the L1 norm, M is the number of pixels and P is a set of the positions of reliable depth estimates, *i.e.*, $P = \left\{p | R^l(p) = 1\right\}$ for pixel coordinate p. Similarly, the distillation loss from StereoNet to MonoNet for the right-view, denoted as L_{dr}, is defined as

$$L_{dr}\left(D^r_m, D^r_s\right) = \frac{1}{M}\left\|D^r_m(Q) - D^r_s(Q)\right\|, \tag{7}$$

where $Q = \{q | R^r(q) = 1\}$ for pixel coordinate q. By using L_{dl} and L_{dr} as the training objective for the finetuning of MonoNet, we can make MonoNet produce more reliable depth estimates.

Several loss terms widely used for depth estimation are also found to be helpful for our finetuning of MonoNet. First, the reconstruction loss L_{re} [45] between I^l and $\tilde{I}^l$ is defined as follows:

$$L_{re}\left(I^l, \tilde{I}^l\right) = \alpha \frac{1\text{–}SSIM\left(I^l, \tilde{I}^l\right)}{2} + \frac{(1-\alpha)}{M}\left\|I^l - \tilde{I}^l\right\|, \tag{8}$$

where *SSIM* measures the structural similarity [48], α is a weighting factor. Second, the depth smoothness loss L_{sm} [45] is given as

$$L_{sm}\left(D^l_m\right) = \frac{1}{M}\sum_p \left\{ \begin{array}{l} \left|\nabla_x D^l_m(p)\right| e^{-\left\|\nabla_x I^l(p)\right\|} \\ + \left|\nabla_y D^l_m(p)\right| e^{-\left\|\nabla_y I^l(p)\right\|} \end{array} \right\}, \tag{9}$$

where ∇_x and ∇_y measure the gradient along the x-axis and y-axis, respectively. Third, the left-right consistency loss L_{lrc} [45] is given as

$$\begin{aligned} L_{lrc}\left(D^l_m, D^r_m\right) &= \frac{1}{M}\left\|D^l_m(P) - \tilde{D}^l_m(P)\right\|, \\ \tilde{D}^l_m &= W\left(D^r_m, D^l_m\right). \end{aligned} \tag{10}$$

Note that L_{re} and L_{sm} are measured only for the left-view, and L_{lrc} is measured for both the left-view and the right-view.

Because MonoNet and StereoNet output multi-scale predictions, each loss term is measured at multi-scales as $L_* = \sum_{n=0}^{N-1} w_n L^n_*$, $* \in \{dl, dr, re, sm, lrc\}$. Here, L^n_* measures each loss for the n-th scale image, and we chose $N = 4$ with the weighting factors as $w_1 = 1.0$, $w_2 = 0.5$, $w_3 = 0.1$, and $w_4 = 0.01$.

IV. EXPERIMENTAL RESULTS

A. DATASETS

Due to the lack of ground-truth depth maps of old photos, we used the SceneFlow and KITTI datasets for the pretraining of MonoNet and StereoNet. Then, for the finetuning of MonoNet, we constructed an old photo dataset by collecting photos from the repository [1].

1) SceneFlow

[43] is a synthetic dataset, containing more than 39,000 stereo pairs for training and 4,000 for testing. Following [35], we obtained the occlusion masks by applying the left-right consistency check and excluded the occluded pixels during the training.

2) KITTI

Reference [47] is a collection of images captured from vehicles in several outdoor scenes. We used the raw data, which include rectified stereo sequences, calibration information, and 3D LIDAR point clouds. The ground-truth depth maps

FIGURE 5. Samples of old photos.

were obtained by mapping the LIDAR points to the image coordinates. In particular, the Eigen split [27] of the KITTI dataset, which contains 22,600, 888, and 687 image pairs for training, validation, and testing, respectively, was used in our experiments.

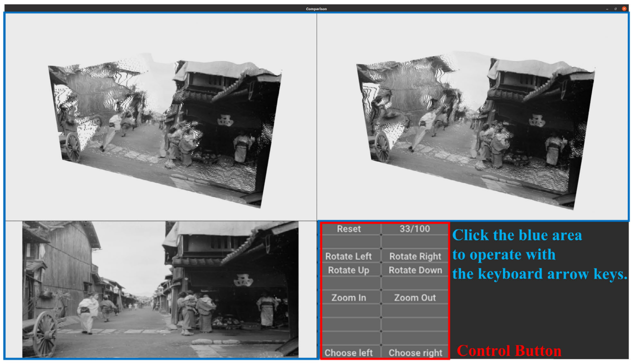
3) OLD PHOTO DATASET

is a collection of old photos. We crawled old photos from the online library [1]. The five collections (*i.e.*, Carpenter, Abdul Hamid II, Lawrence & Houseworth, Grabill, Travel view of Japan and Korea) were selected out of 70 collections in the Print & Photographs online catalog. From the selected collections, we extracted 100 photos in consideration of various types of scenes to define our old photo dataset. Fig. 5 shows old photo examples. Our subject quality evaluation was conducted on only 20 images to reduce the evaluation time, but the results for all images are available at the project page.[1] These old photos were taken around the 1900's with various cameras and resolutions. The old photos were boundary-cropped and pre-processed [2] to be used as input for depth estimation networks.

B. IMPLEMENTATION DETAILS

1) StereoNet

was pretrained using the SceneFlow and KITTI datasets. We adopted DispNetC [43] for the StereoNet architecture but modified the output layers to obtain depth maps for both left-view and right-view to enable the proposed finetuning of MonoNet for each old photo. StereoNet was first trained for 50 epochs on the SceneFlow dataset with a batch size of 4 and a patch size of 768 × 384. The initial learning rate was 10^{-4} and scaled by half at 20, 35, and 45 epochs. StereoNet was then finetuned on the KITTI dataset with a patch size of 832 × 256. The learning rate was initialized as 2×10^{-5} and decayed until it reached 2.5×10^{-6}. For the subsequent unsupervised finetuning, StereoNet was trained for 10 epochs on the Eigen split of the KITTI dataset.



Click the blue area to operate with the keyboard arrow keys.

FIGURE 6. Pairwise comparison evaluation tool.

2) MonoNet

was pretrained using the KITTI dataset. We adopted VGG-16 [44] for the MonoNet architecture but modified the output layers to obtain both left-view and right-view depth maps. MonoNet was pretrained for 50 epochs with a batch size of 4. The input images and depth maps from StereoNet were resized to 512 × 256 to fit the input size of MonoNet. The learning rate was initialized as 10^{-4} and decayed by half at 20, 35, and 45 epochs. The encoder of MonoNet was initialized with the ImageNet-pretrained model.

3) OUR PROPOSED METHOD

adopted zero-shot learning, thus finetuned MonoNet for 10 epochs with the learning rate of 10^{-6} for each given old photo. Note that the parameters of StereoNet were fixed during the finetuning of MonoNet. As for the rendering of stereo images, the warping module from Monodepth [45] was used. The input size for MonoNet and StereoNet was set as 512 × 256 for training.

C. PERFORMANCE EVALUATION

Because ground-truth depth maps do not exist for old photos, we conducted subjective quality evaluation. In addition, several objective quality metrics that do not require ground-truth were used to quantitatively compare the performance. Full 32-bit floating-point precision was used for all trained and tested models.

1) OBJECTIVE QUALITY EVALUATION

typically requires ground-truth data for comparison. However, conventional measures, such as RMSE and the percentage of bad pixels, are not applicable to old photos due to the lack of the ground-truth. Other no-reference quality measures developed for color images, such as NIQE [53], PI [54], and NIMA [55], are also not suitable for the quality evaluation of depth images. We thus first evaluated the performance

[1] https://github.com/rmawngh/Old-Photo-3D

TABLE 1. Performance comparisons with different quality metrics. ↓ and ↑ represent the lower the better and the higher the better, respectively.

	RR-DQM [49] (↓)	NIQSV [50] (↑)	NIQSV+ [51] (↑)	MNSS [52] (↑)
MonoNet	0.1838	26.2759	6.7699	0.0265
Proposed	**0.0784**	**26.3104**	**6.7825**	**0.0399**
CADepth [9]	0.2033	26.3044	6.7743	0.0004
HRDepth [10]	0.8125	26.1094	6.7227	0.0029
MonoDepth2 [11]	0.1996	26.3042	6.7676	0.0008

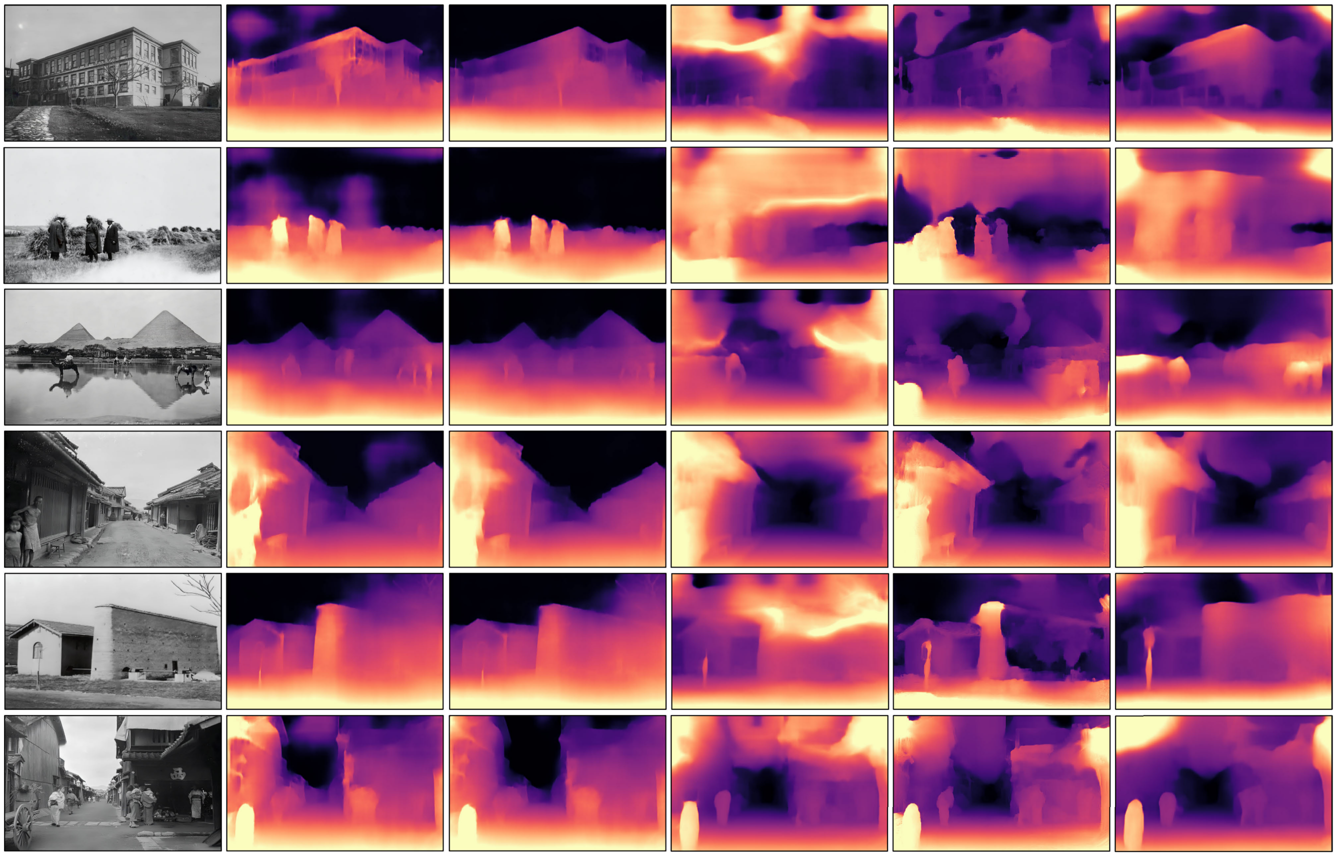

(a) Old Photos (b) MonoNet (c) Proposed Method (d) CADepth-Net (e) HRDepth (f) MonoDepth2

FIGURE 7. Qualitative performance comparisons.

of depth estimation methods using RR-DQM [49] that is a depth quality assessment metric requiring only a pair of color and depth images. In particular, RR-DQM measures local image distortions caused by image rendering with the estimated depth map. We compared the proposed method with state-of-the-art monocular depth estimation methods, including CADepthNet [9], HRDepth [10] and MonoDepth2 [11]. Table 1 shows the RR-DQM results for 20 images that are used for the subjective evaluation. The smaller the depth distortion, the lower the RR-DQM score. The results demonstrate that the proposed method outperforms the state-of-the-art methods.

We also evaluated the performance using depth image-based rendering (DIBR) quality assessment metrics [50], [51], [52]. These metrics measure the distortions in DIBR-synthesized images. Among DIBR quality assessment metrics, we used no-reference methods, including NIQSV [50], NIQSV+ [51], and MNSS [52], to compare the proposed method with other state-of-the-art methods. NIQSV [50] uses edges detected by morphological operations for quality measurement, and NIQSV+ [51] further applies stretching detection and black hole detection. MNSS [52] is a metric derived from multi-scale natural scene statistics. The results shown in Table 1 demonstrate that the proposed method outperforms the state-of-the-art methods.

2) SUBJECTIVE QUALITY EVALUATION

was performed in consideration of our objective of enabling a more realistic rendering of old photographs. Specifically, three-dimensional (3-D) renderings of old photos were obtained using the estimated depth maps, and the subjects were asked to assess the quality of the 3-D renderings. Since

TABLE 2. Performance comparisons on the Kitti eigen test split using different loss combinations, where L_d is a distillation loss, L_{re} is a reconstruction loss, L_{sm} is a smoothness loss, and L_{lrc} is a left-right consistency loss.

	Abs Rel (↓)	Sq Rel (↓)	RMS (↓)	Log RMS (↓)	$\delta < 1.25$(↑)	$\delta < 1.25^2$(↑)	$\delta < 1.25^3$(↑)
L_d	0.1023	0.7304	4.4944	0.1853	**0.8753**	0.9613	0.9831
$L_d + L_{re}$	**0.1017**	0.7205	4.4489	0.1829	0.8772	0.9624	0.9836
$L_d + L_{re} + L_{sm}$	0.1019	0.7148	4.4581	0.1821	0.8754	0.9623	0.9838
$L_d + L_{re} + L_{sm} + L_{lrc}$	0.1019	**0.7144**	**4.4293**	**0.1810**	0.8764	**0.9627**	**0.9843**

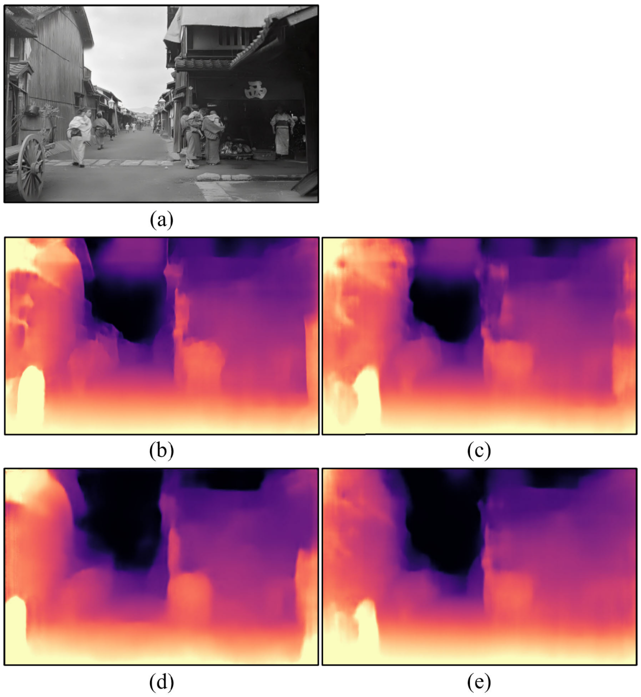


FIGURE 8. Example of the left-view and right-view depth maps obtained by the proposed method: (a) is an original old photo, (b) and (c) are the depth maps obtained using the pretrained MonoNet, and (d) and (e) are the depth maps obtained by the proposed method.

there was no evaluation tool for 3-D rendering comparison, we built it using the Open3D API [56]. As shown in Fig. 6, our evaluation tool consists of two 3-D renderings, a single old photo, and a control panel. We conducted a pairwise comparison by asking the subjects to indicate their preference from two randomly shuffled 3-D renderings. The subjects were recommended to use the keyboard to navigate around the 3-D renderings, and one minute was given for the decision on one image pair.

We compared the proposed method with MonoNet [35], the baseline model that our method is applied. 20 subjects participated in the subjective quality evaluation. The result shows that MonoNet [35] and the proposed method were preferred over the other method by 6.35 and 13.65 times on average out of total 20 times, demonstrating the effectiveness of the proposed method.

Fig. 7 shows several qualitative comparisons with other methods. The proposed method outperforms the other state-of-the-art monocular depth estimation methods [9], [10], [11]. Furthermore, Fig. 8 shows the estimated depth maps for the left and right viewpoints. Compared to MonoNet [35], our proposed method provides enhanced left and right viewpoint depth maps. More results can be found on our project website.

D. ABLATION STUDY

Since the total loss in (5) consists of different loss terms, we investigated the effectiveness of each loss term as ablation studies. For quantitative performance evaluation, we used the ground-truth depth maps from the KITTI dataset [47] for this experiment since old photos do not have ground-truth depth maps. Specifically, each image of the KITTI Eigen split [27] was tested using different models trained with different combinations of the loss terms.

The proposed method trains MonoNet using reliable estimates from StereoNet using the distillation loss (L_d in Table 2), and the other three loss terms, *i.e.*, the reconstruction loss L_{re}, smoothness loss L_{sm}, and left-right consistency loss L_{lrc}, are auxiliary loss terms for regularization. We thus tested different models by including additional loss terms to the distillation loss. The average performance scores for the left-view depth maps of the KITTI Eigen split [27] shown in Table 2 demonstrate that the model trained with all loss terms produced the best performance in most quality metrics, and the other terms contributed to the performance improvements. The weighting factors for the loss terms were empirically chosen as $\mu_1 = 0.5$, $\mu_2 = 0.1$, and $\mu_3 = 0.01$.

For the performance analysis on our target old photos, we provide multiple resultant images obtained using the models trained with different loss configurations on our project page.

V. CONCLUSION

Although old photos have archaeological and historical significance, depth estimation of old photos has attracted more attention. Because most old photos are available as single-view images and their ground-truth depth maps cannot be available, we developed a learning framework that is based on the collaboration of monocular and stereo depth estimation networks. Specifically, the monocular network was used to produce input for the stereo network, and the stereo network was used to yield reliable depth predictions to be used for supervision of the monocular network training. We could train the monocular network to produce a high-quality depth map of the given old photo by the proposed method. From the qualitative and quantitative performance evaluations, we demonstrated the effectiveness of the proposed method.

REFERENCES

[1] Library of Congress. *Prints & Photographs Online Catalog*. Accessed: Jan. 30, 2023. [Online]. Available: https://www.loc.gov/pictures/

[2] Z. Wan, B. Zhang, D. Chen, P. Zhang, D. Chen, J. Liao, and F. Wen, "Bringing old photos back to life," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2747–2757.

[3] V. Bruni and D. Vitulano, "A generalized model for scratch detection," *IEEE Trans. Image Process.*, vol. 13, no. 1, pp. 44–50, Jan. 2004.

[4] I. Giakoumis, N. Nikolaidis, and I. Pitas, "Digital image processing techniques for the detection and removal of cracks in digitized paintings," *IEEE Trans. Image Process.*, vol. 15, no. 1, pp. 178–188, Jan. 2006.

[5] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4471–4480.

[6] S. Zhao, J. Cui, Y. Sheng, Y. Dong, X. Liang, E. I. Chang, and Y. Xu, "Large scale image completion via co-modulated generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–15.

[7] S. Farooq Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth estimation using adaptive bins," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4008–4017.

[8] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," 2019, *arXiv:1907.10326*.

[9] J. Yan, H. Zhao, P. Bu, and Y. Jin, "Channel-wise attention-based network for self-supervised monocular depth estimation," in *Proc. Int. Conf. 3D Vis. (3DV)*, Dec. 2021, pp. 464–473.

[10] X. Lyu, L. Liu, M. Wang, X. Kong, L. Liu, Y. Liu, X. Chen, and Y. Yuan, "HR-depth: High resolution self-supervised monocular depth estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 3, pp. 2294–2301.

[11] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3828–3838.

[12] A. Pizurica, L. Platisa, T. Ruzic, B. Cornelis, A. Dooms, M. Martens, H. Dubois, B. Devolder, M. De Mey, and I. Daubechies, "Digital image processing of the Ghent altarpiece: Supporting the painting's study and conservation treatment," *IEEE Signal Process. Mag.*, vol. 32, no. 4, pp. 112–122, Jul. 2015.

[13] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[14] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for Markov random fields," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 16–29.

[15] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2007.

[16] Q. Yang, L. Wang, R. Yang, H. Stewénius, and D. Nistér, "Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 492–504, Mar. 2009.

[17] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2006, pp. 404–417.

[18] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5410–5418.

[19] Z. Liang, Y. Feng, Y. Guo, H. Liu, W. Chen, L. Qiao, L. Zhou, and J. Zhang, "Learning for disparity estimation through feature constancy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2811–2820.

[20] S. Khamis, S. Fanello, C. Rhemann, A. Kowdle, J. Valentin, and S. Izadi, "StereoNet: Guided hierarchical refinement for real-time edge-aware depth prediction," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 573–590.

[21] I. Cherabier, J. L. Schonberger, M. R. Oswald, M. Pollefeys, and A. Geiger, "Learning priors for semantic 3D reconstruction," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 314–330.

[22] G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia, "SegStereo: Exploiting semantic information for disparity estimation," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 636–651.

[23] G. Botella, A. Garcia, M. Rodriguez-Alvarez, E. Ros, U. Meyer-Baese, and M. C. Molina, "Robust bioinspired architecture for optical-flow computation," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 4, pp. 616–629, Apr. 2010.

[24] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4161–4170.

[25] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8934–8943.

[26] Y. Ming, X. Meng, C. Fan, and H. Yu, "Deep learning for monocular depth estimation: A review," *Neurocomputing*, vol. 438, pp. 14–33, May 2021.

[27] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 2366–2374.

[28] J. Li, R. Klein, and A. Yao, "A two-streamed network for estimating fine-scaled depth maps from single RGB images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3372–3380.

[29] Z. Zhang, C. Xu, J. Yang, J. Gao, and Z. Cui, "Progressive hard-mining network for monocular depth estimation," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3691–3702, Aug. 2018.

[30] X. Yang, Y. Gao, H. Luo, C. Liao, and K.-T. Cheng, "Bayesian DeNet: Monocular depth prediction and frame-wise fusion with synchronized uncertainty," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2701–2713, Nov. 2019.

[31] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. Eur. Conf. Comput. Vis.*, 2017, pp. 740–756.

[32] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1851–1858.

[33] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5667–5675.

[34] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[35] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang, "Learning monocular depth by distilling cross-domain stereo networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 484–500.

[36] A. Pilzer, S. Lathuiliere, N. Sebe, and E. Ricci, "Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9768–9777.

[37] H. Liu, J. Yuan, C. Wang, and J. Chen, "Pseudo supervised monocular depth estimation with teacher–student network," 2021, *arXiv:2110.11545*.

[38] Z. Wu, X. Wu, X. Zhang, S. Wang, and L. Ju, "Learning depth from single image using depth-aware convolution and stereo knowledge," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.

[39] Y. Wang, X. Li, M. Shi, K. Xian, and Z. Cao, "Knowledge distillation for fast and accurate monocular depth estimation on mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2457–2465.

[40] X. Ye, X. Fan, M. Zhang, R. Xu, and W. Zhong, "Unsupervised monocular depth estimation via recursive stereo distillation," *IEEE Trans. Image Process.*, vol. 30, pp. 4492–4504, 2021.

[41] J. Cho, D. Min, Y. Kim, and K. Sohn, "Deep monocular depth estimation leveraging a large-scale outdoor stereo dataset," *Expert Syst. Appl.*, vol. 178, Sep. 2021, Art. no. 114877.

[42] K. Song and K.-J. Yoon, "Learning monocular depth estimation via selective distillation of stereo knowledge," 2022, *arXiv:2205.08668*.

[43] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4040–4048.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[45] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 270–279.

[46] M. Poggi, F. Tosi, and S. Mattoccia, "Learning monocular depth estimation with unsupervised trinocular assumptions," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 324–333.

[47] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[48] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[49] T.-H. Le, S.-W. Jung, and C. S. Won, "A new depth image quality metric using a pair of color and depth images," *Multimedia Tools Appl.*, vol. 76, no. 9, pp. 11285–11303, 2017.

[50] S. Tian, L. Zhang, L. Morin, and O. Deforges, "NIQSV: A no reference image quality assessment metric for 3D synthesized views," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1248–1252.

[51] S. Tian, L. Zhang, L. Morin, and O. Deforges, "NIQSV+: A no-reference synthesized view quality assessment metric," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1652–1664, Apr. 2018.

[52] K. Gu, J. Qiao, S. Lee, H. Liu, W. Lin, and P. Le Callet, "Multiscale natural scene statistical analysis for no-reference quality evaluation of DIBR-synthesized views," *IEEE Trans. Broadcast.*, vol. 66, no. 1, pp. 127–139, Mar. 2019.

[53] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Dec. 2012.

[54] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6228–6237.

[55] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[56] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," 2018, *arXiv:1801.09847*.

JU HO KIM received the B.S. degrees in multimedia engineering from Dongguk University, Seoul, South Korea, in 2015 and 2021. He is currently pursuing the M.S. degree with the Department of Electrical Engineering, Korea University, Seoul. His current research interest includes depth estimation and their applications.

KWANG-LIM KO received the B.S. degrees in mechanical engineering from Sungkyunkwan University, Suwon, South Korea, in 2017 and 2021. He is currently pursuing the M.S. degree with the Department of Automotive Convergence, Korea University, Seoul, South Korea. His current research interests include human pose and shape estimation, depth estimation, and their applications in intelligent transportation.

LE THANH HA (Member, IEEE) received the B.S. and M.S. degrees in information technology from the College of Technology, Vietnam National University, Hanoi, and the Ph.D. degree, in 2010. In 2005, he received a Korean Government Scholarship for the Ph.D. program at the Department of Electronics Engineering, Korea University. After graduation, he joined the Faculty of Information Technology, University of Engineering and Technology—Vietnam National University, as an Associate Professor. His research interests include image/video analysis and processing, satellite image processing, and computer vision. He has deep experiences in teaching digital image processing, computer vision, and multimedia communication courses for both undergraduate and postgraduate programs. He has also been a principle investigator and a main investigator of many fundamental research and technology development projects funded by both domestic and international organizations. He also makes contributions in serving many domestic and international ICT academic conferences, including KSE, NICS, ATC, SoICT, and ICEIC. In addition, he is a member of The Institute of Electronics, Information and Communication Engineers (IEICE) and The Vietnamese Association for Pattern Recognition (VAPR).

SEUNG-WON JUNG (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2005 and 2011, respectively. He was a Research Professor with the Research Institute of Information and Communication Technology, Korea University, from 2011 to 2012. He was a Research Scientist with the Samsung Advanced Institute of Technology, Yongin, South Korea, from 2012 to 2014. He was an Assistant Professor with the Department of Multimedia Engineering, Dongguk University, Seoul, from 2014 to 2020. In 2020, he joined the Department of Electrical Engineering, Korea University, where he is currently an Associate Professor. He has published over 70 peer-reviewed articles in international journals. His current research interests include image processing and computer vision. He received the Hae-Dong Young Scholar Award from the Institute of Electronics and Information Engineers, in 2019.