Dual CNN Models for Unsupervised Monocular Depth Estimation

Vamshi Krishna Repala and Shiv Ram Dubey[(✉)]

Computer Vision Group, Indian Institute of Information Technology, Sri City,
Chittoor, Andhra Pradesh, India
{vamshi.r14,srdubey}@iiits.in

Abstract. The unsupervised depth estimation is the recent trend by utilizing the binocular stereo images to get rid of depth map ground truth. In unsupervised depth computation, the disparity images are generated by training the CNN with an image reconstruction loss. In this paper, a dual CNN based model is presented for unsupervised depth estimation with 6 losses (DNM6) with individual CNN for each view to generate the corresponding disparity map. The proposed dual CNN model is also extended with 12 losses (DNM12) by utilizing the cross disparities. The presented DNM6 and DNM12 models are experimented over KITTI driving and Cityscapes urban database and compared with the recent state-of-the-art result of unsupervised depth estimation.

Keywords: Dual CNN · Depth estimation · Unsupervised · Deep learning

1 Introduction

The image based depth estimation of scene is a very active research area in the field of computer vision. The depth map from images can be estimated in various ways like structure from motion [14], multi-view stereo [19], monocular methods [17], single-image methods [18], etc. The deep learning and convolutional neural networks (CNNs) based methods perform outstanding in most of the problems of computer vision such as image classification [10], facial micro-expression recognition [15], face anti-spoofing [13], hyper-spectral image classification [16], image-to-image transformation [9], colon cancer nuclei classification [1], etc. Inspired from the success of deep learning, several researchers also tried to utilize the CNN for the depth prediction, specially in monocular imaging conditions. These approaches are classified mainly in three categories namely learning-based stereo [21,23], supervised single view depth estimation [3,11], and unsupervised depth estimation [4,6]. The stereo image pairs and ground truth disparity data are needed in order to train the learning-based stereo models. In real scenario, creating such data is very difficult. Moreover, these methods generally create the artificial data which can not represent the real challenges appearing in natural images and depth maps. The supervised single view depth estimation methods

also use ground truth depth to train the model. The main hurdle in supervised approaches is availability and creation of ground truth depth maps which is always not available in real applications.

The unsupervised depth estimation methods do not need any ground truth depth maps. Basically, they utilize the underlying theory of epipolar constraints [7]. Recently, Garg et al. used auto-encoder deep CNN to predict the inverse depth map (i.e. disparity) from left image [4]. They computed a warp image (i.e. reconstructed left image) from disparity map and right image. Finally, the error between original and reconstructed left image is used as the loss to train the whole setup in unsupervised manner. This approach is further improved by Godard et al. by incorporating the left-right consistency [6]. In left-right consistency, basically two depth maps (i.e. left and right) are generated using auto-encoder only from the left input image. The left input image is used with generated right depth map and the right image is used with generated left depth map to reconstruct the right and left images respectively. Zhou et al. [22] utilized the concepts of unsupervised image depth estimation proposed in [3] and [6] to tackle the monocular depth and camera motion estimation in unstructured video sequences in unsupervised learning framework. In one of the recent work, the 3D loss such as photometric quality of frame reconstructions is combined with 2D loss such as pixel-wise or gradient-based loss for learning the depth and ego-motion from monocular video in unsupervised manner [12].

While the unsupervised based methods have gained the attention in recent times, there is still need of discovering better suited unsupervised networks and loss functions. Through this paper, we propose a dual CNN based model for unsupervised monocular image depth estimation by utilizing the 6 losses (DNM6). We also extend the dual CNN model with 12 losses and generate DNM12 architecture to improve the quality of depth maps. The appearance matching loss, disparity smoothness loss and left-right consistency loss are used in this paper. The rest of the paper is structured by presenting the proposed dual CNN models DNM6 and DNM12 in Sect. 2, the experimental results and analysis in Sect. 3, and the concluding remarks in Sect. 4.

2   Proposed Methodology

2.1   Dual Network Model with 6 Losses (DNM6)

The proposed idea of dual network model (DNM) using CNN is illustrated in Fig. 1. This model is based on the 6 losses, thus referred as the DNM6 model. The DNM6 model has two CNN one for each left and right images of stereo pair. During training, the left image I^l and right image I^r are considered as the inputs to the left CNN named as CNN-L and right CNN named as CNN-R respectively. The $I_{i,j}$ refers to the $(i, j)^{th}$ co-ordinate of image I. It is assumed that both I^l and I^r images are captured in similar settings. Both CNN's are based on the auto-encoder algorithm and combined these two networks named as dual network. The CNN architecture (in both CNNs) is taken from the Godard et al. [6]. The CNN-L predicts the left disparity map d^l, whereas the CNN-R
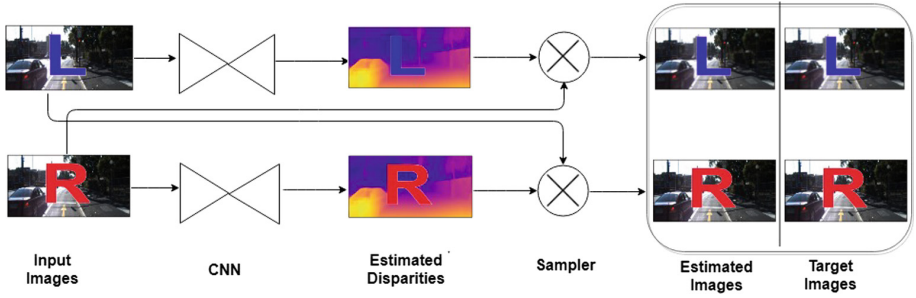
Fig. 1. Pictorial representation of proposed dual network model with 6 losses (DNM6)

predicts the right disparity map d^r. The $d_{i,j}$ refers to disparity value at $(i,j)^{th}$ co-ordinate of disparity map d. In order to reconstruct the left and right image from left and right disparity maps (d^l and d^r), the bilinear sampling from the Spatial Transform Networks [8] is used in this paper. The similar approach is also followed in [6] for reconstruction from disparity map. The left image is reconstructed from the left disparity map d^l and input right image I^r, whereas the right image is reconstructed from the right disparity map d^r and input left image I^l as shown in the Fig. 1. The reconstructed left and right images are referred as $\hat{I}^l$ and $\hat{I}^r$ respectively throughout the paper. We also used the loss functions (C) such as appearance matching loss (C_{ap}), disparity smoothness loss (C_{ds}) and left-right consistency loss (C_{lr}) similar to [6] but in dual network framework. The loss functions are defined below.

Appearance Matching loss: To enforce the appearance of estimated images must be similar to the input image, a combination of L1 norm and Structural Similarity Index Metric (SSIM) [20] loss term is used for both left and right images, defined as [6],

$$C_{ap}^{\beta} = \frac{1}{N} \sum_{i,j} \alpha \frac{1 - SSIM(I_{ij}^{\beta}, \hat{I}_{ij}^{\beta})}{2} + (1 - \alpha) \parallel I_{ij}^{\beta} - \hat{I}_{ij}^{\beta} \parallel \tag{1}$$

where $\beta \in \{l, r\}$, C_{ap}^l refers appearance matching loss between estimated left image and input left image and C_{ap}^r refers appearance matching loss between estimated right image and input right image and α represents the weight between SSIM and L1 norm.

Disparity Smoothness Loss: The image gradient based disparity smoothness loss is computed from both disparity maps to ensure the estimated disparity map should be smooth. Similar to [6], the disparity smoothness loss is given as,

$$C_{ds}^{\beta} = \frac{1}{N} \sum_{i,j} |\partial_x d_{ij}^{\beta}| e^{-\|\partial_x I_{ij}^{\beta}\|} + |\partial_y d_{ij}^{\beta}| e^{-\|\partial_y I_{ij}^{\beta}\|} \tag{2}$$
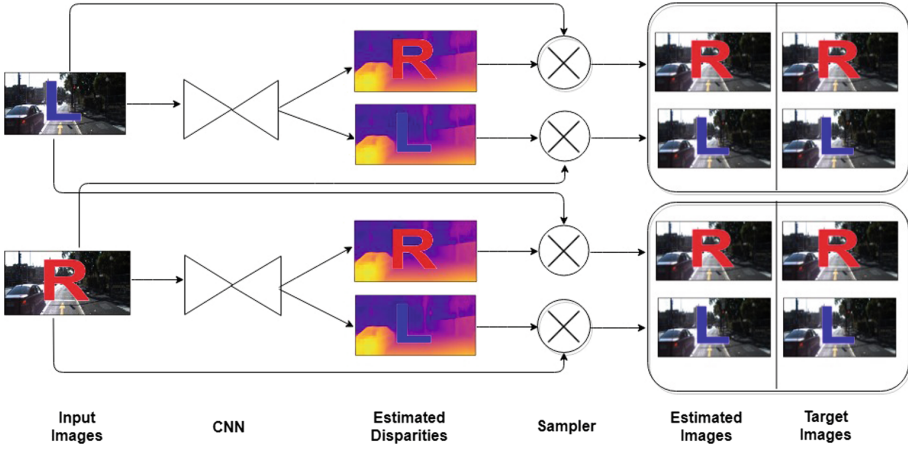
Fig. 2. Pictorial representation of our dual network model with 12 losses (DNM12)

where $\beta \in \{l, r\}$, C_{ds}^l refers the disparity smoothness loss of left disparity map d^l estimated by CNN-L, C_{ds}^r refers the disparity smoothness loss of right disparity map d^r estimated by CNN-R and ∂ is the partial derivative.

Left Right Consistency Loss: To maintain the estimated left disparity map d^l and right disparity map d^r to be consistent, the L1 term penalties on estimated disparities similar to [6] are computed between d^l and d^r as follows,

$$C_{lr} = \frac{1}{N} \sum_{i,j} |d_{ij}^l - d_{ij+d_{ij}^l}^r| \quad \text{and} \quad C_{rl} = \frac{1}{N} \sum_{i,j} |d_{ij}^r - d_{ij+d_{ij}^r}^l| \tag{3}$$

where C_{lr} and C_{rl} refer the left to right and right to left consistency losses respectively.

Similar to Godard et al. [6], four output scales s in both left and right CNNs are used in this paper in order to make the loss functions more robust. The combined cost function C_s at scale s including all above losses i.e. appearance matching losses C_{ap}^l and C_{ap}^r, disparity smoothness losses C_{ds}^l and C_{ds}^r and left-right consistency losses C_{lr} and C_{rl} is given as $C_s = \alpha_{ap}(C_{ap}^l + C_{ap}^r) + \alpha_{ds}(C_{ds}^l + C_{ds}^r) + \alpha_{lr}(C_{lr} + C_{rl})$. The final Cost/Loss function for proposed DNM6 model is computed as $C = \sum_{s=1}^4 C_s$ at different output scales from s = 1 to 4 similar to [6]. At testing time, a single left image, I^l is needed as the input to the left CNN (i.e., CNN-L) and it predicts the disparity map d^l from the trained network. Note that, the right CNN with input I^r can also be used to predict the disparity map d^r. Once disparity map d (i.e. d^l or d^r) is computed, it can be converted into depth map (D) as $D = \frac{f \times B}{d}$, where f represents the focal length and B is the baseline between stereo cameras.

2.2 Dual Network Model with 12 Losses (DNM12)

In our previous DNM6 model, disparity maps are estimated from each network individually, whereas in this DNM12 model, the left-right cross disparity mapping is also proposed as depicted in Fig. 2. The left and right CNN networks of DNM6 are extended to generate two output disparities (i.e. left and right) from each CNN. Similar to Godard et al. [6], it generates both left and right disparity maps from a single image. During training, the left image I^l and right image I^r of stereo pair are provided as inputs to the left CNN (CNN-L) and right CNN (CNN-R) respectively. In DNM12 architecture, both the CNN's predict the left and right disparities independently as illustrated in Fig. 2. Here, we consider d^{l_l} and d^{l_r} as the left and right disparity maps respectively estimated by the left CNN-L and similarly d^{r_l} and d^{r_r} as the left and right disparity maps respectively estimated by the right CNN-R. As shown in the Fig. 2, four bilinear samplers are used for reconstructing the two output left images $\hat{I}^{l_l}$ and $\hat{I}^{r_l}$ corresponding to left input image and two output right images $\hat{I}^{l_r}$ and $\hat{I}^{r_r}$ corresponding to right input image. The $\hat{I}^{l_l}$ uses d^{l_l} and I^r, $\hat{I}^{l_r}$ uses d^{l_r} and I^l, $\hat{I}^{r_l}$ uses d^{r_l} and I^r, and $\hat{I}^{r_r}$ uses d^{r_r} and I^l. In DNM12, four appearance matching losses, four disparity smoothness losses and four left-right consistency losses are considered.

The *Four Appearance Matching Losses* are defined as follows,

$$C_{ap}^{\beta_\gamma} = \frac{1}{N} \sum_{i,j} \alpha \frac{1 - SSIM(I_{ij}^\gamma, \hat{I}_{ij}^{\beta_\gamma})}{2} + (1 - \alpha) \parallel I_{ij}^\gamma - \hat{I}_{ij}^{\beta_\gamma} \parallel \qquad (4)$$

where $\beta \in \{l, r\}$, $\gamma \in \{l, r\}$, $C_{ap}^{l_l}$ and $C_{ap}^{l_r}$ are the appearance matching losses for left CNN-L and $C_{ap}^{r_l}$, $C_{ap}^{r_r}$ are the appearance matching losses for right CNN-R. The total appearance matching loss is given by $C_{ap} = (C_{ap}^{l_l} + C_{ap}^{l_r} + C_{ap}^{r_l} + C_{ap}^{r_r})$.

The *Four Disparity Smoothness Losses* are computed as follows,

$$C_{ds}^{\beta_\gamma} = \frac{1}{N} \sum_{i,j} |\partial_x d_{ij}^{\beta_\gamma}| e^{-\|\partial_x I_{ij}^\beta\|} + |\partial_y d_{ij}^{\beta_\gamma}| e^{-\|\partial_y I_{ij}^\beta\|} \qquad (5)$$

where $\beta \in \{l, r\}$, $\gamma \in \{l, r\}$, $C_{ds}^{l_l}$, $C_{ds}^{l_r}$ are the disparity smoothness losses for left CNN-L and $C_{ds}^{r_l}$, $C_{ds}^{r_r}$ are the disparity smoothness losses for right CNN-R. The total disparity smoothness loss is computed as $C_{ds} = (C_{ds}^{l_l} + C_{ds}^{l_r} + C_{ds}^{r_l} + C_{ds}^{r_r})$.

The *Four Left-Right Consistency Losses* are calculated as follows,

$$C_{lr}^l = \frac{1}{N} \sum_{i,j} |d_{ij}^{l_l} - d_{ij+d_{ij}^{l_l}}^{l_r}|, \quad \text{and} \quad C_{rl}^l = \frac{1}{N} \sum_{i,j} |d_{ij}^{l_r} - d_{ij+d_{ij}^{l_r}}^{l_l}| \qquad (6)$$

$$C_{lr}^r = \frac{1}{N} \sum_{i,j} |d_{ij}^{r_l} - d_{ij+d_{ij}^{r_l}}^{r_r}|, \quad \text{and} \quad C_{rl}^r = \frac{1}{N} \sum_{i,j} |d_{ij}^{r_r} - d_{ij+d_{ij}^{r_r}}^{r_l}| \qquad (7)$$

where C_{lr}^l, C_{rl}^l are the left-right and right-left consistency losses for left CNN-L and C_{lr}^r, C_{rl}^r are the left-right and right-left consistency losses for right CNN-R. The total left-right consistency loss is calculated as $C_{lr} = (C_{lr}^l + C_{rl}^l + C_{lr}^r + C_{rl}^r)$.

Similar to DNM6, the total Loss function in DNM12 is also defined as $C = \sum_{s=1}^{4} C_s$ at different output scales from s = 1 to 4, where C_s at a particular scale is computed by weighted sum of all losses as $C_s = \alpha_{ap} \times C_{ap} + \alpha_{ds} \times C_{ds} + \alpha_{lr} \times C_{lr}$. The same procedure as provided in previous DNM6 model is followed in DNM12 also for testing, a single image is taken as input to either CNN-L or CNN-R and it predicts the disparity map from the trained network which is converted into depth map.

Table 1. Experimental results by using proposed dual CNN based DNM6 and DNM12 models for unsupervised depth estimation over KITTI benchmark database. The training is done over KITTI training images and the evaluation is done over KITTI test images. In this table, pp denotes the post-processing. The best results without post-processing are highlighted in bold face.

Method	Lower is better					Higher is better		
	Abs Rel	Sq Rel	RMSE	RMSE log	d1-all	a1	a2	a3
Godard et al. [6] No LR	0.123	1.417	6.315	0.220	30.318	0.841	0.937	0.973
Godard et al. [6]	0.124	1.388	6.125	0.217	**30.272**	0.841	0.936	0.975
DNM6 Model	0.1223	1.4004	6.162	0.214	31.050	**0.848**	**0.941**	**0.976**
DNM12 Model	**0.1221**	**1.3058**	**6.069**	**0.213**	31.455	0.841	0.939	**0.976**
DNM6 Model PP	0.1157	1.2037	5.830	0.203	30.004	0.852	0.945	0.979
DNM12 Model PP	0.1157	1.1404	5.772	0.203	30.342	0.848	0.944	0.979

3   Experimental Results and Analysis

We have used the standard datasets such as KITTI and Cityscapes for the experiments. The KITTI database [5] consists of stereo pairs from different scenes. Similar to Godard's work [6], 29,000 stereo pairs are used for training and 200 high-quality images are used as the test cases along with its depth maps. The Cityscapes database [2] contains the stereo pairs captured for autonomous driving. Similar to Godard's work [6], we have used the 22,973 stereo pairs for training after cropping each image such that the 80% of the height is preserved and the car hoods are removed. Similar to [6], we have used the same 200 KITTI stereo images for testing over Cityscapes database.

The CNN architectures in our network are same as in Godard et al. [6]. The proposed DNM6 and DNM12 models are implemented in TensorFlow which contains 62 million trainable parameters. We have used following parameters, $\alpha = 0.85$, $\alpha_{ap} = 1$, $\alpha_{ds} = 0.1$, $\alpha_{lr} = 1.0$ and learning rate $\lambda = 10^{-4}$ for first 30 epochs and 0.5 x 10^{-4} for next 10 epochs and 0.25 x 10^{-4} for the last 10 epochs. The data augmentation is done on fly, similar to [6]. During test time, a post-processing is performed to reduce the effect of stereo dis-occlusions similar to [6].

In both DNM6 and DNM12 methods, the estimated disparity map $d(x)$ is further converted into depth map as $D(x) = \frac{fB}{d(x)}$, where f is the focal length and B is the baseline. The evaluation of both models are done with the estimated depth maps $D(x)$ and provided ground truth depth maps $G(x)$. The evaluation metrics are same as in [6] such as Absolute Relative difference (**Abs Rel**), Squared Relative difference (**Sq Rel**), Root Mean Square Error (**RMSE**), **RMSE log**, and **d1-all**. The lower values of these metrics represent the better performance. We also measured the *Accuracy metrics* (i.e., **a1**, **a2**, and **a3** similar to [6]) for which higher is better.

The results are reported in Table 1 over KITTI database and compared with very recent state-of-the-art unsupervised method proposed by Godard et al. [6] with and without left-right (LR) consistency. Note that the lower values of **Abs Rel**, **Sq Rel**, **RMSE**, **RMSE log**, and **d1-all** and the higher values of accuracies **a1**, **a2**, and **a3** represent the better performance. The performance of proposed DNM6 and DNM12 methods are also tested with a pre-processing (**PP**) step to reduce the effect of stereo dis-occlusions [6]. The best results without PP are highlighted in bold face in Table 1. It can be easily observed that the proposed dual CNN based models i.e. both DNM6 and DNM12 perform better than Godard et al. [6] with and without left-right consistency. The **Abs Rel**, **Sq Rel**, **RMSE**, **RMSE log**, and **d1-all** values are generally lower and accuracies **a1**, **a2**, and **a3** are higher for the proposed DNM6 and DNM12 methods. It is also noticed that DNM12 completely outperforms the Godard et al. [6] in all terms except **d1-all**. The performance of DNM6 model is improved in terms of the **Abs Rel**, **RMSE**, **a1**, **a2**, and **a3** as compared to the Godard model. The DNM12 model exhibits the better performance as compared to the DNM6 model in all terms except accuracies. As for as accuracies are concerned, the DNM6 model is superior as compared to DNM12 model because generating right disparity from left image and left disparity from right image is not suited for pixel level thresholding. This is also seen that the performance of proposed models improved significantly with post-processing step over KITTI database.

Table 2. Experimental results by using proposed dual CNN based DNM6 and DNM12 models for unsupervised depth estimation over Cityscapes benchmark database. The training is done over Cityscapes training images and the evaluation is done over KITTI test images. In this table, pp denotes the post-processing. The best results without post-processing are highlighted in bold face.

Method	Lower is better					Higher is better		
	Abs Rel	Sq Rel	RMSE	RMSE log	d1-all	a1	a2	a3
Godard et al. [6]	0.699	10.060	14.445	0.542	94.757	0.053	0.326	0.862
DNM6 Model	0.2704	3.7637	9.186	0.326	64.215	0.649	0.864	0.941
DNM12 Model	**0.2661**	**3.6491**	**8.915**	**0.316**	**61.163**	**0.669**	**0.875**	**0.946**
DNM6 Model PP	0.2474	2.9781	8.406	0.300	63.780	0.663	0.881	0.954
DNM12 Model PP	0.2396	2.8945	8.178	0.289	58.733	0.687	0.889	0.959

The results comparison of proposed models with Godard et al. [6] over Cityscapes database is illustrated in Table 2. In this Table, the training is performed over Cityscapes database, whereas the test images are same as in KITTI database. It is noticed from this experiment that the proposed models are superior than Godard et al. [6] over Cityscapes database in all terms. Moreover, the DNM12 model performs better than DNM6 model. As for as both databases are concerned, the results of proposed models over KITTI database is better than the Cityscapes database. The possible reason can be the difference between the camera calibration between training and testing databases. The similar observations are also made by Godard et al. [6]. The post-processing step enhances the performance of proposed DNM6 and DNM12 models over Cityscapes database.

4   Conclusion

In this paper, the dual CNN based models DNM6 and DNM12 are presented for unsupervised monocular depth estimation. The dual network models used two different CNNs (CNN-L and CNN-R) for left and right images of training stereo pairs respectively. In DNM6 and DNM12, total 6 and 12 losses are used, respectively. The results are computed over benchmark KITTI and Cityscapes databases and compared with the recent left-right consistency based method. It is observed that the DNM12 outperforms the existing method left-right consistency method. It is also observed that the DNM12 model improves the performance over DNM6 model in most of the cases. The post-processing step further boosts the performance of proposed models.

References

1. Basha, S.S., Ghosh, S., Babu, K.K., Dubey, S.R., Pulabaigari, V., Mukherjee, S.: RCCNet: an efficient convolutional neural network for histological routine colon cancer nuclei classification. In: IEEE ICARCV, pp. 1222–1227. IEEE (2018)
2. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: IEEE CVPR, pp. 3213–3223 (2016)
3. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NIPS, pp. 2366–2374 (2014)
4. Garg, R., BG, V.K., Carneiro, G., Reid, I.: Unsupervised CNN for single view depth estimation: geometry to the rescue. In: ECCV, pp. 740–756 (2016)
5. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: IEEE CVPR, pp. 3354–3361 (2012)
6. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: IEEE CVPR (2017)
7. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge (2003)
8. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: NIPS, pp. 2017–2025 (2015)

9. Kancharagunta, K.B., Dubey, S.R.: CSGAN: cyclic-synthesized generative adversarial networks for image-to-image transformation. arXiv preprint arXiv:1901.03554 (2019)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)
11. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. IEEE TPAMI **38**(10), 2024–2039 (2016)
12. Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints. arXiv:1802.05522 (2018)
13. Nagpal, C., Dubey, S.R.: A performance evaluation of convolutional neural networks for face anti spoofing. arXiv preprint arXiv:1805.04176 (2018)
14. Nistér, D.: Preemptive ransac for live structure and motion estimation. Mach. Vis. Appl. **16**(5), 321–329 (2005)
15. Reddy, S.P.T., Karri, S.T., Dubey, S.R., Mukherjee, S.: Spontaneous facial micro-expression recognition using 3D spatiotemporal convolutional neural networks. arXiv preprint arXiv:1904.01390 (2019)
16. Roy, S.K., Krishna, G., Dubey, S.R., Chaudhuri, B.B.: HybridSN: exploring 3D–2D CNN feature hierarchy for hyperspectral image classification. arXiv preprint arXiv:1902.06701 (2019)
17. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: NIPS, pp. 1161–1168 (2006)
18. Saxena, A., Chung, S.H., Ng, A.Y.: 3-D depth reconstruction from a single still image. IJCV **76**(1), 53–69 (2008)
19. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. IEEE CVPR **1**, 519–528 (2006)
20. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004)
21. Zbontar, J., LeCun, Y.: Stereo matching by training a convolutional neural network to compare image patches. J. Mach. Learn. Res. **17**(1–32), 2 (2016)
22. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: IEEE CVPR (2017)
23. Zhou, T., Krahenbuhl, P., Aubry, M., Huang, Q., Efros, A.A.: Learning dense correspondence via 3D-guided cycle consistency. In: IEEE CVPR, pp. 117–126 (2016)