Automatic Depth Estimation from Single 2D Image via Transfer Learning Approach

Muhammad Awais Shoukat[1], Allah Bux Sargano*[1], Zulfiqar Habib[1], and Lihua You[2]

[1]Department of Computer Science, COMSATS University Islamabad, Lahore, Pakistan
[2]National Centre for Computer Animation, Bournemouth University, United Kingdom

***Abstract*—Nowadays, depth estimation from a single 2D image is a prominent task due to its numerous applications such as 2D to 3D image/video conversion, robot vision, and self-driving cars. This research proposes an automatic novel technique for the depth estimation of single 2D images via transfer learning of pre-trained deep learning model. This is a challenging problem, as a single 2D image does not carry any cues regarding depth. To tackle this, the pool of available images is exploited for which the depth is known. By following the hypothesis that the color images having similar semantics are most probably to have similar depth. Along these lines, the depth of the input image is predicted through corresponding depth maps of semantically similar images available in the dataset, fetched by high-level features of pre-trained deep learning model followed by a classifier (i.e., K-Nearest Neighbor). Afterward, a Cross Bilateral filter is applied for the removal of fallacious depth variations in the depth map. To prove the quality of the presented approach, different experiments have been conducted on two publicly available benchmark datasets, NYU (*v*2) and Make3D. The results indicate that the proposed approach outperforms state of the art methods.**

***Index Terms*—Depth estimation, 2D to 3D conversion, transfer learning, KNN-Framework.**

I. Introduction

When conventional 2D cameras capture the pictures, the depth information is lost as a result of projection of the scene onto the 2D image plane. However, to estimate this lost dimension for recovering the 3D structure has become of more importance in these years. Furthermore, a huge increment has occurred in the accessibility of 3D Players (3D TVs, cinemas, smart-phones, and projectors etc), but the volume of 3D content has still not attained the growth to that extent. To overcome such issues, a number of advanced techniques have been cited in the literature for converting 2D images/videos to 3D. Generally, depth estimation followed by Depth-Image-Based Rendering are the key steps involved in the conversion of 2D contents to 3D [1]. This paper only focuses on depth estimation from the single 2D image. Estimating the depth can enable numerous applications such as 2D image/video to 3D conversion [2], robot vision for navigation in space [3], it can also be used for classifying and recognizing the images [4].

There are two prominent strategies to deal with depth estimation of single 2D image i.e., semi-automatic and automatic. The semi-automatic approach allows the human's interference, for assigning sparse depths to various positions of the scene. Then, these assigned depth values are propagated through the entire scene to generate a dense depth map [1]. Involvement of human interference makes these methods very costly and time-consuming. In an automatic approach, generally, no human interference is allowed. Based on automatic approach, various algorithms have been proposed in the literature for the estimation of depth using different cues like defocus, motion and shading. Most of these approaches are rely on some heuristic assumptions and specific settings. For example, 'structure from motion' algorithm assumes that the camera is moving continuously, and shape from shading and texture based techniques [5]–[8], rely on the uniform color and texture. However, these algorithms can perform well in simple and confined scenarios [1], but do not perform well on complex images.

Recently, an alternative to heuristics-based depth estimation approaches, a reasonable assumption based on the presence of a relationship between the visual appearance of images and its depth values has come under the attention. These techniques are based on the hypothesis that the color images having photometric similarity are most probably to have similar 3D structure. Motivated by this observation, several data-driven approaches have been developed. In such approaches, a pool of available images in the dataset (image + depth) having similarity with the input image is retrieved. The matching process in these algorithms is typically based on different kinds of features. For depth estimation of the single image, a feature descriptor GIST over saliency map is used to asses the similarity between dataset images and given input image in [9]. Another handcrafted feature descriptor Histogram of Oriented Gradients (HOG) is applied to directly fetch photo-metrically similar images to the input image in [10]. Afterward, the depth map is generated by fusing the depth maps of retrieved images followed by a Cross-Bilateral filter to enrich the estimated depth map. Recently, a model has been proposed, where the author extracted similar images to the input image by embedding (texture, blurriness, color) features with relative height [3]. A sampling approach followed by Markov Random Field is proposed in [11]. An adaptive approach to select the variable number of similar images using Local Binary Patterns (LBP) features is proposed in [12]. Following the same methodology, a combination of most frequently used feature descriptors (GIST, LBP, HOG, and SURF) is used for better results in [1]. In this model, the author first categorized the images into the different clusters after sampling. Then, the

*Corresponding author: allahbux@cuilahore.edu.pk

978-1-7281-1929-8/18/$31.00 ©2018 IEEE
DOI 10.1109/EECS.2018.00114

best-matched cluster images to the input image are selected on the basis of the proposed combination of feature descriptors.

In practice, the whole depth map of an input image is predicted through the corresponding depth maps of photo-metrically similar images extracted through the different combination of hand-crafted features. However, each descriptor usually performs good for some specific types of images, and there is no universal feature descriptor available that performs best on every type of images. The selected combination of feature descriptor may perform well in a dataset and lack performance on other. So, it's not a viable option to select a specific combination of feature descriptor. To overcome this limitation in most of the object classification and recognition domains, researchers focus has been shifted from handcrafted feature extraction to deep learning based techniques. However, high computational resources and the large amount of data is required for training of deep learning model from scratch. In the literature, some of the authors proposed the data-augmentation technique to enlarge the dataset for the suitable training of deep learning model [13], [14]. This is not a good approach as it increases training computation and may not be suitable for the real-time scenario. Hence, training of deep learning model from scratch is not an appropriate approach for the domain-specific problems [15], where the size of the dataset is small. On the other hand, some recent studies in image recognition and classification tasks used the concept of transfer learning (domain adaptation), fine-tune the deeply learned models on a specific task to a new task even in a changed domain [15]–[18]. The transfer learning is also favorable for limited size dataset training and can also be used for real-time applications. There are two ways to approach transfer learning:

1) Fine-tune the weights based on the target dataset, while preserving the original pre-trained network.

2) Use of pre-trained network for only feature extraction and categorize those features through a suitable classifier.

Based on the second option, we proposed a strategy for automatically learning of relevant features directly from input data through transfer learning approach using pre-trained deep learning model i.e., Residual Neural Network (ResNet-50 [19]). ResNet-50 competed the ImageNet 'Large Scale Visual Recognition Challenge' in 2015 with the error rate of 3.57%. There are also other publicly available models i.e., AlexNet [20] and GoogLeNet [21], but due to their high error rate ResNet-50 [19] model is used. The idea has been taken from the approach adopted in human action recognition by using transfer learning with deep representations [15]. Where the author used AlexNet [20] as a feature extractor, afterward, an ensemble classifier is used to recognize the human actions. In our proposed solution, after extraction of high-level features from pre-trained deep learning model ResNet-50, K-Nearest Neighbor framework (KNN) is used for the appropriate selection of photo-metrically alike images available in the training set. Afterward, the whole depth map is generated by fusing depth maps of those selected images. The complete algorithm has been discussed in the Proposed Methodology section. The experiments are conducted on two public and widely-used data-sets NYU ($v2$) [22] and Make3D [11], to prove the efficiency and effectiveness of the algorithm.

II. PROPOSED METHODOLOGY

The proposed automatic depth estimation algorithm can be illustrated as follows. Provided an input image, and an RGB-D dataset (Make3D or NYU $v2$), comprised of RGB images whose depth information is available, the aim is to estimate the depth map of the input image. Fig. 1 also demonstrates the whole process involved in the proposed methodology. Following are the major steps involved:

1) Sampling of the input image into 4x4 tiles to preserve the positions of objects. This approach will help to distinguish the geometry of the scene.

2) Feature extraction of the sampled image using pre-trained deep learning model i.e., ResNet-50. The feature vector F_i representing i^{th} image is extracted by the concatenating the features of every tile of i^{th} image as shown in

$$F_i = [f_{i,1} \; f_{i,2} \; f_{i,3} \; \cdots \; f_{i,n}], \tag{1}$$

where $f_{i,t==1}$ represents the feature vector with t^{th} tile number and i^{th} image.

3) Find photo-metrically similar images to the input image, available in the dataset using high-level features extracted in the previous step followed by K-Nearest Neighbor framework with correlation as a similarity metric and number of nearest neighbors *k*=4.

4) Analysis of correlation values of fetched images to remove outliers [12]. Only those images will be selected whose correlation values greater than a pre-defined correlation coefficient (Threshold=0.50). Fig. 2 and Fig. 3 shows some examples of selected images.

5) Depth map generation of the input image by fusing the corresponding depth maps of selected structurally similar images. Fig. 4 shows some examples of generated depth maps by our proposed algorithm. If D is the combination of the depth maps, $C_{val}[i]$ correlation value of i^{th} image and $D[i]$ associated depth map of i^{th} image. The fusion process is illustrated as

$$D = \frac{1}{\sum_i C_{val}[i]} \left[\sum_i C_{val}[i] \times D[i] \right]. \tag{2}$$

6) Finally, the refinement on the generated depth map has been done through the Cross-Bilateral-Filter to remove fallacious depth variations in the depth map.

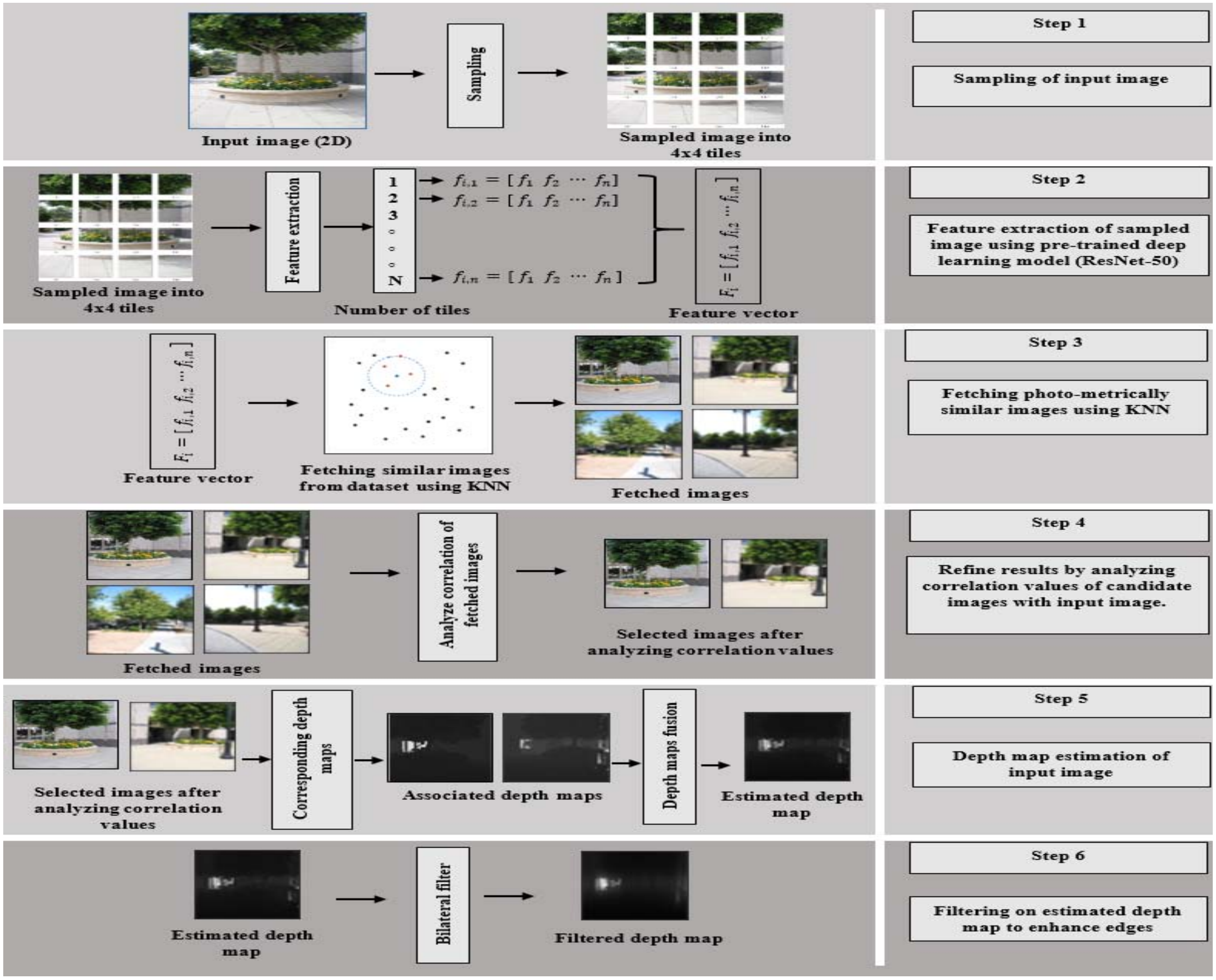
Input image (2D)
Sampling
Sampled image into 4x4 tiles
Step 1
Sampling of input image
Sampled image into 4x4 tiles
Feature extraction
1
2
3
N
$f_{i,1} = [f_1\ f_2\ \cdots\ f_n]$
$f_{i,2} = [f_1\ f_2\ \cdots\ f_n]$
$f_{i,n} = [f_1\ f_2\ \cdots\ f_n]$
Number of tiles
$F_i = [f_{i,1}\ f_{i,2}\ \cdots f_{i,n}]$
Feature vector
Step 2
Feature extraction of sampled image using pre-trained deep learning model (ResNet-50)
$F_i = [f_{i,1}\ f_{i,2}\ \cdots f_{i,n}]$
Feature vector
Fetching similar images from dataset using KNN
Fetched images
Step 3
Fetching photo-metrically similar images using KNN
Fetched images
Analyze correlation of fetched images
Selected images after analyzing correlation values
Step 4
Refine results by analyzing correlation values of candidate images with input image.
Selected images after analyzing correlation values
Corresponding depth maps
Associated depth maps
Depth maps fusion
Estimated depth map
Step 5
Depth map estimation of input image
Estimated depth map
Bilateral filter
Filtered depth map
Step 6
Filtering on estimated depth map to enhance edges

Fig. 1. Block diagram of proposed system.

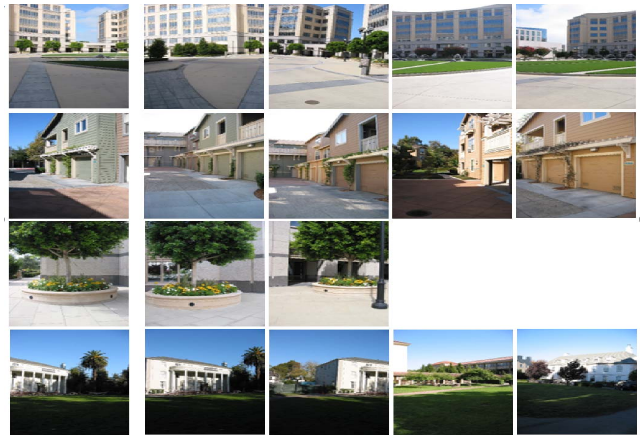

Fig. 2. 2D query image (Column-1) and variable number of nearest neighbors (Columns 2-5) retrieved using feature extraction via transfer learning followed by KNN classifier on Make3D dataset.

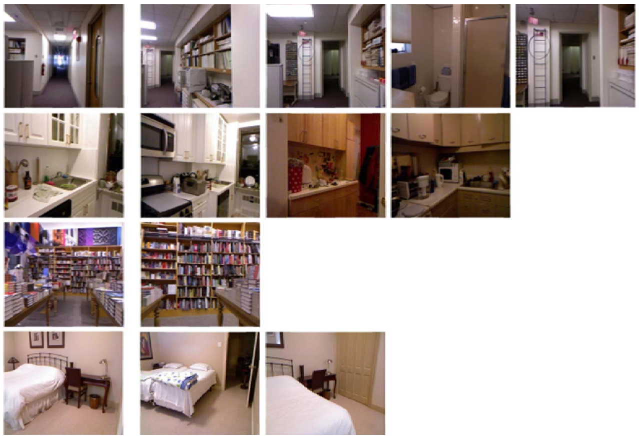

Fig. 3. 2D query image (Column-1) and variable number of nearest neighbors (Columns 2-5) retrieved using feature extraction via transfer learning followed by KNN classifier on NYU ($v2$) dataset.

III. EXPERIMENTATIONS AND RESULTS

To assess the quality of the proposed algorithm, the experiments have been conducted on two different datasets (Make3D and NYU (*v*2)). The NYU data set is comprised 1449 RGB images (795 for training set and 654 for test set) along with their corresponding depth information of indoor scenes. While Make3D dataset contains 534 (400 for training set and 134 for test set) outdoor scenes. For more experiments on NYU dataset, leave one out methodology is used for comparison with state of the art. This selects one by one (image + depth) pair from the dataset as a test set, while leaving all the other pairs for the training set. In the literature, some authors also combined these two datasets to generate a new dataset and reported the results on it. To compare with those models, the combination of both two datasets is used to generate a new dataset with 1983 images and leave one out methodology is used for comparison. Following measures are used to evaluate the quality of the proposed algorithm. If P is the number of pixels, M estimated depth map, M^* ground truth, μ_M, μ_{M^*} statistical mean values, σ_M and σ_{M^*} statistical standard deviation of estimated and ground truth depth maps respectively, the generalized mathematical form of the error measures can be derived as

$$C = \frac{\sum_i (M[i] - \mu_M)(M^*[i] - \mu_{M^*})}{P \ \ \sigma_M \ \ \sigma_{M^*}}, \qquad (3)$$

$$log10 = \frac{1}{P}\sum_i [\log_{10}(M^*[i]) - \log_{10}(M[i])], \qquad (4)$$

$$RMSE = \sqrt{\frac{\sum_i (M^*[i] - M[i])^2}{P}}, \qquad (5)$$

$$PSNR = 20\log_{10}\frac{Max(M^*)}{RMSE}, \qquad (6)$$

$$RMSE(log) = \sqrt{\frac{\sum_i (log_{10}(M^*[i]) - log_{10}(M[i]))^2}{P}}, \qquad (7)$$

$$REL = \frac{1}{P}\sum_i \left[\frac{|M^*[i] - M[i]|}{M^*[i]}\right]. \qquad (8)$$

For comparison of our algorithm with state of the art, above mentioned error measures have been taken. These error rates are used to check the quality of the estimated depth map generated by our algorithm with actual ground truth depth map. Table 1-4 shows the comparison results of the proposed method with state of the art. The results of the other models have been taken from their publications or cited by other papers. The proposed approach shows improved results than state of the art algorithms. The best results have been represented through bold values. The sign '-' indicates that the result of the cited paper is not available for that particular measure. On NYU dataset, our algorithm outperforms the other state of the art algorithms, while on Make3D dataset the results are close to the state of the art. Higher the value of C, and PSNR represents high quality, while lower the results of REL, Log10, RMSE, and RMSE log is better. The proposed approach falls into the category of algorithms where results depend on the selection of depth-wise similar images. With high-level feature more accurate candidates are selected, which reduces the error rate. Hence, better quality results are achieved through this novel approach. Fig. 4 shows some examples of generated depth maps by our proposed algorithm.

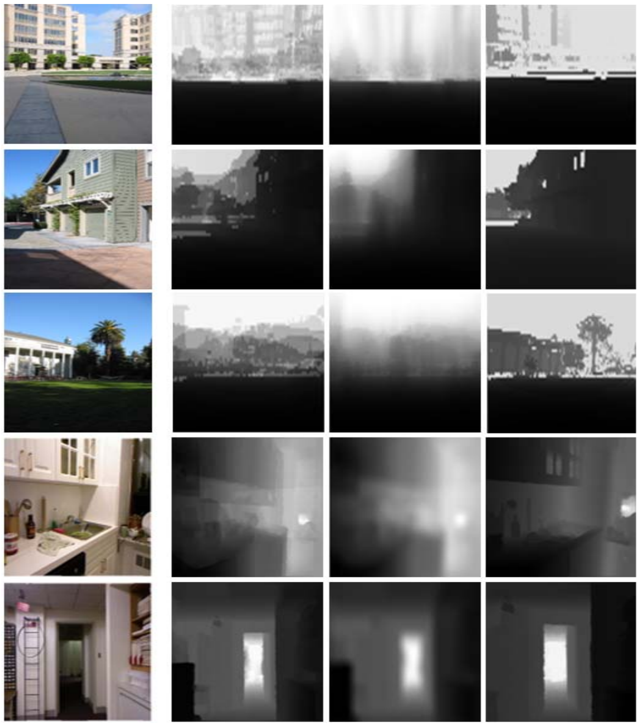

Fig. 4. 2D query image, estimated depth by proposed algorithm, refined depth map and actual ground truth depth map (Columns 1-4) respectively.

IV. CONCLUSION

This research work proposed a novel depth estimation algorithm using transfer learning technique. A strategy for automatically learning of relevant features directly from input data using pre-trained deep learning model ResNet-50 has been introduced for the estimation of depth. The algorithm relies on the hypothesis that the color images having photometric similarity likely present similar depth structure. Based on this idea, algorithm finds structurally similar images of input image from the pool of available images whose depth is known. Then, the depth map of the input image is estimated through the fusion of corresponding depth maps of those structurally alike images. In the last step, a Cross Bilateral filter is used to remove the fallacious depth variations in the estimated depth map. Different accuracy measures are used to assess the quality of the presented algorithm. The algorithm achieved better/closer results to the most of the renowned algorithms in state-of-the-art. In our future directions, we will improve the mechanism of selection of more relevant images to reduce more error rate. Combination of pre-trained deep learning models can also be tried.

TABLE I

Comparison of state of the art algorithms on NYU ($v2$) dataset. The sign '-' indicates that the result of the listed paper is not available for that particular measure. Best results are represented through bold characters.

NYU - Train-795 and Test-654 Split				
Algorithm	**RMSE**	**RMSE(log)**	**REL**	**Log10**
Proposed	0.956	**0.152**	0.325	**0.124**
TRNN [3]	1.12	0.402	0.395	0.144
Depth transfer [23]	1.2	-	0.35	0.131
Weighted median [24]	1.28	0.90	1.3	0.29
Make3D [11]	1.214	0.409	0.35	-
Perception in the wild [25]	1.10	0.38	0.34	-
Mid level vision [26]	1.20	0.42	0.40	-
Multi-scale deep network [27]	**0.907**	0.285	**0.215**	-

TABLE II

Comparison of state of the art algorithms on Make3D dataset. The sign '-' indicates that the result of the listed paper is not available for that particular measure. Best results are represented by bold characters.

Make3D - Train-400 and Test-134 split					
Algorithm	**C**	**PSNR**	**RMSE**	**REL**	**Log10**
Proposed	0.623	14.27	15.92	0.49	0.16
Multiple features [1]	0.66	14.10	-	0.407	**0.140**
TRNN [3]	**0.74**	-	**13.43**	0.407	0.145
DEPT [28]	-	-	16.7	0.421	0.172
Adaptive LBP Based [12]	0.66	14.06	14	0.384	0.156
Weighted median [24]	0.66	-	15.94	0.376	0.161
HOG feature based [10]	0.61	13.4	-	0.432	0.18
Batra et al. [29]	-	-	15.8	**0.362**	0.168
Depth Transfer [23]	0.69	**14.56**	15.1	0.362	0.148
Semantic label [30]	-	-	-	0.379	0.148
Make3D [11]	0.64	-	-	0.458	0.149

TABLE III

Comparison of state of the art algorithms on NYU ($v2$) dataset with Leave One Out strategy. Best results are represented by bold characters.

NYU-Leave One Out				
Algorithm	**C**	**PSNR**	**REL**	**Log10**
Proposed	0.62	**15.1**	**0.31**	**0.11**
Multiple features [1]	**0.63**	13.90	0.407	0.140
Adaptive LBP Based [12]	0.63	13.74	0.422	0.153
HOG feature based [10]	0.61	12.90	0.539	0.183
Depth Transfer [23]	0.59	13.57	0.374	0.134

TABLE IV

Comparison of state of the art algorithms on the combination of both two datasets, NYU ($v2$) and Make3D. Best results are represented by bold characters.

NYU and Make3D-Leave One Out				
Algorithm	**C**	**PSNR**	**REL**	**Log10**
Proposed	0.60	**15.31**	**0.388**	**0.134**
Multiple features [1]	**0.63**	14.00	0.41	0.16
Adaptive LBP Based [12]	0.62	13.81	0.568	0.192
Depth Transfer [23]	0.60	12.48	0.559	0.196
HOG feature based [10]	0.60	13.58	1.01	0.245

Acknowledgement

This research is supported by the PDE-GIR project which has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 778035.

References

[1] J. L. Herrera, C. R. del Blanco, and N. García, "Automatic depth extraction from 2d images using a cluster-based learning framework," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3288–3299, 2018.

[2] Y. Wang, R. Wang, and Q. Dai, "A parametric model for describing the correlation between single color images and depth maps," *IEEE Signal Processing Letters*, vol. 21, no. 7, pp. 800–803, 2014.

[3] H. Mohaghegh, N. Karimi, S. R. Soroushmehr, S. Samavi, and K. Najarian, "Aggregation of rich depth aware features in a modified stacked generalization model for single image depth estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[4] A. Torralba and A. Oliva, "Depth estimation from image structure," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 9, pp. 1226–1238, 2002.

[5] T. Lindeberg and J. Garding, "Shape from texture from a multi-scale perspective," in *Computer Vision, 1993. Proceedings., Fourth International Conference on*, pp. 683–691, IEEE, 1993.

[6] J. Malik and R. Rosenholtz, "Computing local surface orientation and shape from texture for curved surfaces," *International journal of computer vision*, vol. 23, no. 2, pp. 149–168, 1997.

[7] A. Maki, M. Watanabe, and C. Wiles, "Geotensity: Combining motion and lighting for 3d surface reconstruction," *International Journal of Computer Vision*, vol. 48, no. 2, pp. 75–90, 2002.

[8] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, "Shape-from-shading: a survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 21, no. 8, pp. 690–706, 1999.

[9] J. L. Arciniegas Herrera, J. Konrad, C. R. d. Blanco Adán, and N. García Santos, "Learning-based depth estimation from 2d images using gist and saliency," in *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 4753–4757, IEEE, 2015.

[10] J. Konrad, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee, "Learning-based, automatic 2d-to-3d image and video conversion," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3485–3496, 2013.

[11] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 824–840, 2009.

[12] J. L. Herrera, C. R. del Bianco, and N. García, "Learning 3d structure from 2d images using lbp features," in *Image Processing (ICIP), 2014 IEEE International Conference on*, pp. 2022–2025, IEEE, 2014.

[13] Y. Kim, H. Jung, D. Min, and K. Sohn, "Deep monocular depth estimation via integration of global and local predictions," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4131–4144, 2018.

[14] L. He, G. Wang, and Z. Hu, "Learning depth from single images with deep neural network embedding focal length," *IEEE Transactions on Image Processing*, 2018.

[15] A. B. Sargano, X. Wang, P. Angelov, and Z. Habib, "Human action recognition using transfer learning with deep representations," in *Neural Networks (IJCNN), 2017 International Joint Conference on*, pp. 463–469, IEEE, 2017.

[16] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, pp. 1520–1528, 2015.

[17] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4293–4302, 2016.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[22] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from rgbd images,” in *European Conference on Computer Vision*, pp. 746–760, Springer, 2012.

[23] K. Karsch, C. Liu, and S. B. Kang, “Depth transfer: Depth extraction from video using non-parametric sampling,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 11, pp. 2144–2158, 2014.

[24] Y. Kim, S. Choi, and K. Sohn, “Data-driven single image depth estimation using weighted median statistics,” in *Image Processing (ICIP), 2014 IEEE International Conference on*, pp. 3808–3812, IEEE, 2014.

[25] W. Chen, Z. Fu, D. Yang, and J. Deng, “Single-image depth perception in the wild,” in *Advances in Neural Information Processing Systems*, pp. 730–738, 2016.

[26] D. Zoran, P. Isola, D. Krishnan, and W. T. Freeman, “Learning ordinal relationships for mid-level vision,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 388–396, 2015.

[27] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in *Advances in neural information processing systems*, pp. 2366–2374, 2014.

[28] H. Qin, X. Li, Y. Wang, Y. Zhang, and Q. Dai, “Depth estimation by parameter transfer with a lightweight model for single still images,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 748–759, 2017.

[29] D. Batra and A. Saxena, “Learning the right model: Efficient max-margin learning in laplacian crfs,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2136–2143, IEEE, 2012.

[30] B. Liu, S. Gould, and D. Koller, “Single image depth estimation from predicted semantic labels,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 1253–1260, IEEE, 2010.