SERBIAN JOURNAL OF ELECTRICAL ENGINEERING
Vol. 17, No. 3, October 2020, 285-296

UDC: 004.89+528.7]:611/612 **DOI: https://doi.org/10.2298/SJEE2003285G**

Automatic Corrections of Human Body Depth Maps using Deep Neural Networks

Gorana Gojić[1], Radovan Turović[1], Dinu Dragan[1], Dušan Gajić[1], Veljko Petrović[1]

Abstract: This paper presents an approach to correcting misclassified pixels in depth maps representing parts of the human body. A misclassified pixel is a pixel of a depth map which, incorrectly, has the ‘background’ value and does not accurately reflect the distance from the sensor to the body being scanned. A completely automatic, deep learning based solution for depth map correction is proposed. As an input, the solution requires a color image and a corresponding erroneous depth map. The input color image is segmented using deep neural network for human body segmentation. The extracted segments are further used as guidance to find and amend the misclassified pixels on the depth map using a simple average based filter. Unlike other depth map refinement solutions, this paper designs a method for the improvement of the input depth map in terms of completeness instead of precision. The proposed method does not exclude the application of other refinement methods. Instead, it can be used as the first step in a depth map enhancement pipeline to determine approximate depths for erroneous pixels, while other refinement methods can be applied in a second step to improve the accuracy of the recovered depths.

Keywords: Depth Map Refinement, Human Body, Photogrammetry, Deep Learning, Segmentation.

1 Introduction

Digital recreation of real-life objects as three-dimensional (3D) models in the virtual digital domain is a topic of growing research and commercial interest. There are many possible ways to acquire data required for the reconstruction of 3D models, including image capturing using a digital camera or mobile device. The sequence of overlapping images captured from different viewpoints can be used in a photogrammetry-based pipeline to obtain a 3D reconstruction from an image sequence. This method is of particular interest because of the exceptionally large number of 2D image acquisition devices

[1]Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, Novi Sad, Serbia;
E-mails: gorana.gojic@uns.ac.rs; radovan.turovic@uns.ac.rs; dinud@uns.ac.rs; dusan.gajic@uns.ac.rs; pveljko@uns.ac.rs

available, in particular, compared to 3D scanners, depth-sensors and the like. The whole pipeline is realized in three phases [1]: Structure from Motion (SfM), Multi-View Stereo (MVS), and mesh creation. The SfM phase uses an overlapping image sequence to produce a sparse point cloud along with camera intrinsic and extrinsic parameters. The output of the SfM phase is used as an input to the MVS phase. As a result, this phase outputs a dense point cloud representing the points of the 3D model's surface in 3D space. A final meshing step uses a dense point cloud to generate a 3D model of the object shown in the image sequence. The completeness of the reconstructed 3D model is highly dependent on the quality of a dense point cloud that MVS phase outputs. The parts of the dense point cloud that are too sparse have a significant chance of not being reconstructed at all. This makes MVS output improvement one of the top priority tasks in terms of 2D to 3D photogrammetry-based reconstruction. Recent benchmarks indicate that there is room for improvement even for the state-of-the-art MVS algorithms, especially in terms of the output completeness [2, 3].

Since the MVS phase is a complex, multi-step process, the aims to improve its output, can focus on improving one or more of the steps that this complex process consists of. This paper concerns itself principally with the completeness of the output of the MVS phase and how it may be improved through the improvement of the depth map estimation process. This estimation is frequently performed as part of many MVS algorithms [4].

A depth map is best conceptualized as a 2D image which complements a traditional colored 2D image with an additional channel of information of the same resolution which has a distance-value for every pixel representing the distance of the particular point in space which is imaged by that pixel. In the case of images forming a photogrammetry set these points are typically assumed to be either part of the background or part of the *surface* of the object being reconstructed. This, naturally, makes the quality of the depth map vitally important to the reconstruction of the object through a mesh of its surface.

It is a well-established experimental fact that MVS algorithms have difficulties with depth estimation if input images are textureless or contain repeated textures [2, 4, 5]. Consequently, misclassified depth pixels, assigned with background depth instead of foreground f++ depth, can be present on the estimated depth map. In this paper, these misclassified values will be referred to as the 'missing depths,' and will represent an error in reconstruction where true depth-value for the values is lost and can only be estimated and/or reconstructed. The study presented in this paper is performed in the context of 2D to 3D photogrammetry-based reconstruction of a clothed human body. This scenario is often plagued with depth-map pixel misclassification precisely because commonly worn clothing patterns present many surfaces which are featureless (such as a plain white shirt) or have repeating textures (any patterned cloth). Since they are caused by a local lack in variation, these misclassifications are usually grouped, covering larger areas, such as whole arms, legs, or the entire back. Since these groups appear on a visualized depth map (and on the reconstructed geometry) as black voids, the terms misclassified group and hole will be used as equivalents in the rest of the paper.

This research is motivated by the observation that human body dense point clouds based on depth maps with problems such as those described are not suitable for further reconstruction, being altogether too sparse for conversion into a mesh. This is most commonly manifested, due to the robustness of the mesh-creation algorithms, as a 3D model of the human body missing one or more body-parts.

To solve the problem, the proposed solution addresses the improvement of depth map completeness as a way to indirectly improve the density of dense point cloud. Therefore, this paper proposes a novel, completely automatic solution for the recovery of missing depths from a depth map showing human subject using machine learning methods. While other solutions for depth map improvement primarily focus on edge improvements on the existing depth map, the proposed solution is able to recover missing depth values from the existing ones under certain conditions which as suggested by the result of empirical research, occur in normal use of such software for commercial or research purposes. As input, the solution requires an RGB image of the subject and the corresponding unrefined depth map which may contain holes. To recover missing depths, a deep neural network is utilized to analyze the RGB image and recognize the human body depicted on it, segmenting it into sections corresponding to major body parts. Incorrectly assigned depths are, then, recovered on a per-segment basis using correctly estimated depth values inside the segment. In other words, the central assumption is that the depth of missing pixels which form, say, the left arm will be most correlated with other left-arm pixels whose depths are known. To further enhance edges and geometry of the recovered depth map, the output from the proposed method can be used as an input to other depth map refinement solutions.

The rest of the paper is organized as follows: In Section 2 a brief overview of the existing depth map refinement methods is presented. Section 3 discusses in detail the proposed method for depth map correction. The obtained results are presented and discussed in Section 4. The final, fifth, section offers the main conclusions, as well as possible directions for future work.

2 Related Work

Many filtering algorithms have been proposed to refine depth maps. Those filters are designed primarily to enhance edges while correcting artifacts on depth maps. In [6] a bilateral filter is proposed. The filter preserves edges while assigning the pixels weighted average of its neighboring pixels. Multiple modifications of bilateral filter, known as join bilateral filter are proposed in [7] and [8]. Unlike the bilateral filter which uses filter input to calculate weights, these filters use another guidance image for weight calculation. Consequently, they overcome over-blur and under-blur issues introduced by the original bilateral filter which, due to its simple method of functioning, smoothed the entire image more or less equally. In [9] a guided image filter has been proposed. In comparison to bilateral and joint bilateral filters, this filter type exhibits superior performance near the edges. The bilateral filter was the basis for the construction of the trilateral filter [10]. Unlike the bilateral filter, this filter provides better noise reduction and better outlier rejection. Multiple modifications in form of joint trilateral filters have been proposed in [1−13]. Whereas the majority of filters refine existing depths by smoothing while preserving edges, [11] is designed to also fill the holes in depth maps. Here, color images are used jointly with depth maps to find and extract misclassified pixels from a depth map, which is then followed with a hole filling step. The authors demonstrate that the method can successfully fill smaller holes on the depth map, but there is no evidence that whole body regions such as arms, or legs can be recovered in this way. Papers [14, 15] utilize the idea of using multiple temporally or spatially connected depth maps for pixel depth correction. In [15] voting-based filter built on the top of the joint bilateral filter is used on a sequence of spatially and temporally connected depth maps. This approach has proven to be successful in correcting depths based on neighboring depth maps. While arguably quite effective, this filter is significantly limited due to the requirement for multiple depth maps taken in temporal sequence. This makes the filter entirely unsuitable to any problem where no more than one point in time is considered or where there are no multiple depth map readings. A depth map refinement solution specific for MVS pipeline is proposed in [14]. The authors propose Z\ VHjopp` a method for outlier reduction based on redundancy. Each pixel is assigned with multiple depths according to other depth maps showing the same point. Then, a Random Markov Field model is used to choose the most probable depth among all depths assigned to the pixel.

Recently, deep neural networks have been successfully applied for the task of depth map estimation and refinement [5, 16 − 19]. The pure depth map refinement solution presented in [5] reduces the noise and boosts geometric details for the input depth map, but does not solve the problem of missing depth values. More often, depth map refinement follows depth map estimation [16, 17, 19]. In those cases, an initial depth map is estimated and then refined by some of the earlier mentioned filters or machine learning techniques. The authors of [17] state that their deep neural network based solution for depth map estimation and refinement even outperforms state-of-the-art photogrammetry software such as Colmap [20, 21].

However, a big problem with neural network-based solutions is ground truth depth map data required to train the network. Ground truth depth data is available in publicly available datasets not related to people, such as NYU v2 [22] and Make3D [23]. Some of the top-performing solutions, such as [17] are trained and tested on those datasets. Human based datasets with ground truth depth maps are hard to find and are usually not free to use. In [5], the authors obtain near ground truth depth maps of the human body by reconstructing a 3D model of a human subject and calculating depths based on the model. This approach for ground truth data generation is not possible if the output model is missing parts of the body, which is the problem addressed in this paper. Further, the use of purely synthetic data of necessity obscures some of the finer details of the actual problem such as, for instance, the peculiarities of lens geometry or the noise presented in most 2D image sensors, to name just two ubiquitous complications. Thus, work with a synthetic ground truth faces an uphill battle to demonstrate that its results are applicable to production (rather than academic dataset) images.

To avoid the problem with missing ground truth depth data, the proposed solution does not use deep neural networks to directly predict a refined depth map. Instead, a deep neural network is used to process the corresponding color image and extract data that will be used for depth map refinement. The idea of using a color image to refine the depth map, though not in this manner, is already known and utilized in previous work [11, 17]. Lastly, the paper presents a simple filter for depth map correction based on average depth calculation which is used to fill the missing values and refine the depth map.

3 The Proposed Method

A complete pipeline for the proposed method is shown in Fig. 1. As an input, the method receives an RGB image I showing a subject and a corresponding depth map D in a Portable Network Graphics (PNG), three-channel RGB format scaled so that each color-component is on the 0 to 255 range and white is (255, 255, 255) while black is (0, 0, 0). The RGB image is further used as a basis for human body segmentation. The extracted human body segment information is then used jointly with the input depth map to generate a corrected depth map.

The reader should note that due to privacy concerns we may not use images in a form which permits user identification. This means that in all illustrations in this paper we do not use the original RGB input images, but instead censor them by showing only the silhouette of the subject. This is a modification of the illustrative images only, and in the actual operation of the system unmodified RGB images are needed and used.

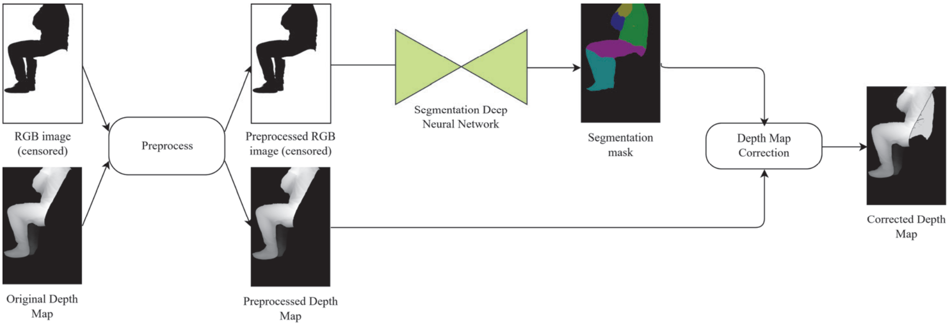

Fig. 1 – *The proposed method pipeline for depth map correction.*

The rest of the section discusses in more detail the proposed method's pipeline. The necessary depth map and RGB image preprocessing techniques are stated. Next, part of the pipeline related to the human body segmentation is explained in detail. Lastly, the method to produce a final depth map is presented.

3.1 Input preprocessing

Since the input RGB image and the corresponding depth map can be of arbitrary size and orientation, both images are first rotated to be of the same orientation. Next, images are scaled to a size which best approximates the optimal size of 480×720 pixels, while keeping the original image size ratio. The purpose of this process is to reduce the size for performance and accuracy reasons without imposing a deforming transformation on the input data.

The target image size is treated as a parameter of the system and, in this case, has been determined empirically through a series of simple experiments. Each experiment is conducted on a set of RGB images showing a human subject. The set is used as a basis to create multiple subsets with images from the original set resized to different dimensions in each subset while keeping the original image ratio to avoid image deformation. The images from the subsets are fed to the pretrained deep neural network model to obtain corresponding segmentation masks. The masks are then compared visually and in terms of accuracy to determine what image dimensions result in optimal segmentation masks. To choose the most accurate mask among multiple segmentation masks, we calculate the average number of pixels classified as a body part for each mask and declare as the optimal the one resulting in the highest average number. The calculation is performed on each mask set and optimal global dimensions are determined by ranking image dimensions according to the accuracy achieved on mask sets for each image and choosing the best ranked dimensions.

3.2 Human body segmentation

This phase of the pipeline uses the preprocessed input RGB image to produce a human body segmentation mask, such as those presented in Fig. 2. A body segmentation mask is yet another image in the vein of the depth map, of the same resolution as the input image, and representing an additional channel of information. This channel, in particular, is one of a set of values which indicate either background or, in the case of foreground, which body segment the corresponding RGB pixel corresponds to.

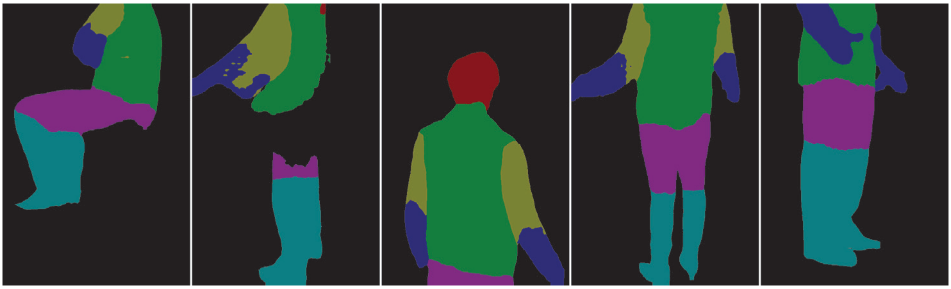

Fig. 2 – *WSHP body segmentation masks.*

For human body segmentation, the latest pretrained model of WSHP human body parsing deep neural network [24] is used to annotate up to seven different segment classes on the input image. Annotations include six body parts: head, torso, left/right upper arms, left/right lower arms, left/right upper legs, left/right lower legs and a special class for the image background. WSHP deep neural network uses DeepLab [25] based shared model proposed in [26] to parse body parts on the input image. DeepLab network is a fully convolutional neural network based on VGG-16 [27] architecture. Contrary to the original VGG-16 architecture, DeepLab replaces the last, fully-connected layer, with a convolutional layer. For more details about the network architecture, a reader is referred to [26]. The segmentation mask produced by the network is further used as an input to the depth map correction phase.

3.3 Depth Map Correction

In the depth map correction phase, the segmentation mask is used jointly with the preprocessed depth map to obtain a corrected depth map. Depth map correction is a two-step process. In the first step, different segments are identified in a segmentation map and average depth color is calculated for each segment. A pseudoalgorithm for this step is shown in Fig. 3. Here, S denotes a segment inside the segmentation mask of size $N\times M$. The erroneous depth map of size $N\times M$ is denoted with D. P represents a pixel at the position (x_P, y_P) in a segmentation mask or depth map. Since both of these images are of the same size, each pixel on a segmentation map has its corresponding pixel at the same coordinates on the depth map and vice-versa. As a result, the first step outputs an average depth value for each segment based on the corresponding depth map. Since the input depth map is represented in a three-channel format, average depth color is calculated per channel (C_R, C_G and C_B in Fig. 3).

A second step is used to assign calculated depth values to the erroneous pixels in the depth map D. All pixels that are classified in a segmentation map as part of the body and represent part of a background in a depth map are assigned with an average depth of a corresponding segment. A pseudoalgorithm for this step is shown in Fig. 4.

```
for all segments (S) in segmentation map do
  if S is not assigned with a background class then
    for all pixels (P) in S do
      if $D(x_P, y_P) \neq$ background depth color then
        $sum_R \leftarrow sum_R + D(x_P, y_P)_R^2$
        $sum_G \leftarrow sum_G + D(x_P, y_P)_G^2$
        $sum_B \leftarrow sum_B + D(x_P, y_P)_B^2$
      end if
    end if
    $N \leftarrow$ number of pixels in segment S
    $C_R \leftarrow \sqrt{sum_R} / N$
    $C_G \leftarrow \sqrt{sum_G} / N$
    $C_B \leftarrow \sqrt{sum_B} / N$
  end if
end for
```

Fig. 3 – *Pseudo-algorithm for segment depth calculation.*

```
for all pixels (P) in erranous depth map D do
  if P is not assigned with a background segment class then
    $S \leftarrow getSegmentForPixel(P)$
    $C \leftarrow getSegmentAverageDepth(\mathrm{S})$
    $D(x_P, y_P)_R \leftarrow C_R$
    $D(x_P, y_P)_G \leftarrow C_G$
    $D(x_P, y_P)_B \leftarrow C_B$
  end if
end for
```

Fig. 4 – *Pseudo-algorithm for pixel depth estimation.*

This means we do not need neighboring pixels to fill holes, allowing for a much greater area of effect. A valid worry, at this stage, is to consider that this averaging may destroy significant geometric detail in the reconstructed object. This can be largely dismissed under the circumstances since if there was significant geometric detail it would cause enough textural variation on the image for the hole not to appear in the first place: holes, as stated above, result principally from flat, uniform, poorly-textured areas which are precisely those for whom the assumption of flatness may be said to hold.

4 Experimental Results

The method is evaluated in the context of photogrammetry-based 2D to 3D reconstruction pipeline. Human body images captured by a multi-view stereo setup are used as input to the AgiSoft MetaShape [28] photogrammetry software, configured for high-quality settings, which produces depth maps. Fig.5 shows the results achieved by the proposed method. As long as a single correct pixel of a segment remains, the whole segment may be reconstructed. The first, third and fourth row of Fig. 5 are examples of completely corrected depth maps, while the second row shows a partially corrected depth map caused by a lack of non-background pixel depths in lower legs segment. Since the recovered segment part is filled with average depth color, the method is particularly successful in recovering segments that are flat, like legs, back or arms. The method is robust to slight vertical viewpoint changes, as shown in the second row of Fig. 5.

To output high-quality depth map correction, the method requires a precisely segmented input image. Poor segmentation can lead to the calculation of incorrect average depth values per segment, as is shown in Fig. 6. On the other hand, if many small segments are detected as part of segmentation this may lead to a lack of valid depth pixels per segment which, in turn, leads to bad average values and, finally, to badly reconstructed holes.

5 Conclusion

In this paper, an end-to-end automatic method for depth map correction in the context of human body reconstruction is presented. The proposed solution addresses the problem of incomplete 3D model reconstruction caused by bad depth map estimation due to surfaces with poor illumination or low texture. It is demonstrated that the proposed method improves such erroneous depth maps by recovering missing values, even when in large groups. This, in turn, improves the reconstructed 3D model. The method relies on the joint use of the unrefined depth map and the corresponding color image with annotated body segments. The best results are achieved under the assumption that human body segmentation is at least moderately precise and that pixels inside the recovered segment are of similar depth. Since the method output is another depth map, the proposed method can be pipelined with other depth map refinement methods to produce both more complete and precise depth maps.

In future work, special attention will be directed towards human body segmentation problem. The method output would benefit from a solution able to segment left and right arm or leg as different segments. Finer body segmentation could also result in better output depth maps since finer segmentation would increase the accuracy of the segment's average depth calculation. To overcome the limitations of the presented method, future research will consider the replacement of average depth calculation step with machine learning methods. This replacement could result in more accurate depths assigned to the recovered depth map pixels.

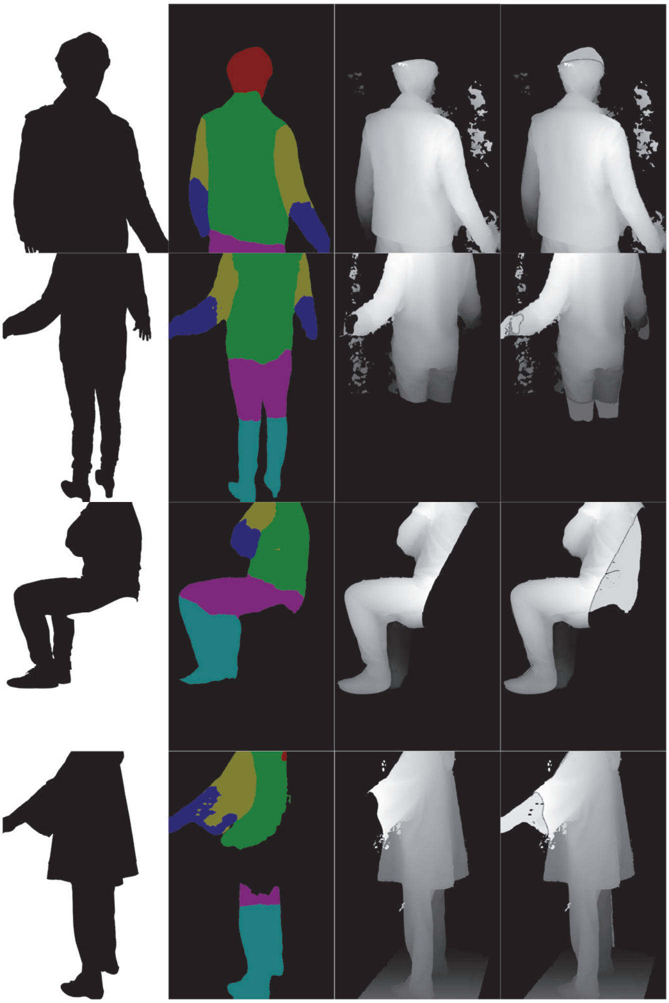

Fig. 5 – *Left to right:* (1) *input image (censored)*;
(2) *segmentation mask produced by WSHP deep neural network*;
(3) *depth map from AgiSoft MetaShape*; (4) *output of the proposed method.*

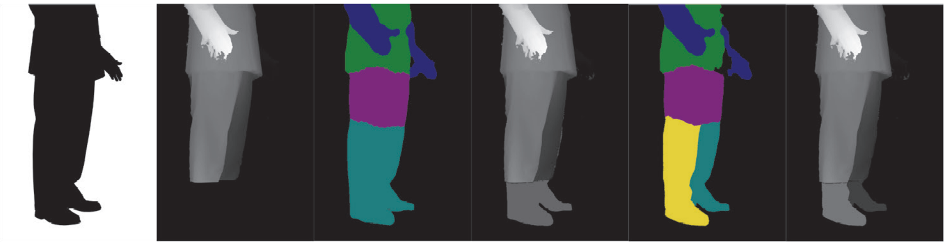

Fig. 6 – *Left to right:* (1) *input image (censored)*; (2) *depth map from AgiSoft MetaShape*; (3) *segmentation map from WSHP neural network*; (4) *output of the proposed method for* (3); (5) *manually corrected segmentation mask*; (6) *output of the proposed method for* (5).

7 References

[1] S. Walker: Close – Range Photogrammetry and 3D Imaging, Photogrammetric Engineering & Remote Sensing, Vol. 81, No. 4, April 2015, pp. 273 – 274.

[2] H. Aanaes, R. R. Jensen, G. Vogiatzis, E. Tola, A. B. Dahl: Large-Scale Data for Multiple-View Stereopsis, International Journal of Computer Vision, Vol. 120, No. 2, April 2016, pp. 153 – 168.

[3] A. Knapitsch, J. Park, Q.- Y. Zhou, V. Koltun: Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction, ACM Transactions on Graphics, Vol. 36, No. 4, July 2017, pp. 78:1 – 78:13.

[4] M. Goesele, B. Curless, S. M. Seitz: Multi-View Stereo Revisited, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, USA, June 2006, pp. 2402 – 2409.

[5] S. Yan, C. Wu, L. Wang, F. Xu, L. An, K. Guo, Y. Liu: DDRNet: Depth Map Denoising and Refinement for Consumer Depth Cameras Using Cascaded CNNs, Proceedings of the 15th European Conferenceon on Computer Vision, Munich, Germany, September 2018, pp. 155 – 171.

[6] C. Tomasi, R. Manduchi: Bilateral Filtering for Gray and Color Images, Proceedings of the 6th International Conference on Computer Vision, Bombay, India, January 1998, pp. 839 – 846.

[7] M. K. Park, J.- H. Cho, I. Y. Jang, S. J. Lee, K. H. Lee: An Iterative Joint Bilateral Filtering for Depth Refinement of a 3D Model, Proceedings of the SA'11 SIGGRAPH Asia 2011 Posters, December 2011, pp. 19:1.

[8] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, K. Toyama: Digital Photography with Flash and No-Flash Image Pairs, ACM Transactions on Graphics, Vol. 23, No. 3, August 2004, pp. 664 – 672.

[9] K. He, J. Sun, X. Tang: Guided Image Filtering, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35, No. 6, June 2013, pp. 1397 – 1409.

[10] P. Choudhury, J. Tumblin: The Trilateral Filter for High Contrast Images and Meshes, Proceedings of the 14th Eurographics Workshop on Rendering Techniques, Leuven, Belgium, June 2003, pp. 186 – 196.

[11] X. Xiang, Z. Yan, C. Nan, W. Xu, L. Zhang: A Modified Joint Trilateral Filter Based Depth Map Refinement Method, Proceedings of the 12th World Congress on Intelligent Control and Automation (WCICA), Guilin, China, June 2016, pp. 1403 – 1407.

[12] S.- W. Jung: Enhancement of Image and Depth Map Using Adaptive Joint Trilateral Filter, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 23, No. 2, February 2013, pp. 258 – 269.

[13] K.- H. Lo, Y.- C. F. Wang, K.- L. Hua: Joint Trilateral Filtering for Depth Map Super-Resolution, Proceedings of the International Conference on Visual Communications and Image Processing, Kuching, Malaysia, November 2013, pp. 1 – 6.

[14] N. D. F. Campbell, G. Vogiatzis, C. Hernández, R. Cipolla: Using Multiple Hypotheses to Improve Depth-Maps for Multi-View Stereo, Proceedings of the 10th European Conference on Computer Vision, Marseille, France, October 2008, pp. 766 – 779.

[15] Y.- H. Chiu, M.- S. Lee, W.- K. Liao: Voting-Based Depth Map Refinement and Propagation for 2D to 3D Conversion, Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference, Hollywood, USA, December 2012, pp. 4377 – 4384.

[16] M. H. Baig, L. Torresani: Coupled Depth Learning, Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Placid, NY, USA, March 2016, pp. 1 – 10.

[17] Y. Yao, Z. Luo, S. Li, T. Fang, L. Quan: MVSNet: Depth Inference for Unstructured Multi-View Stereo, Proceedings of the 15th European Conferenceon on Computer Vision, Munich, Germany, September 2018, pp. 767 – 783.

[18] Y. Cao, Z. Wu, C. Shen: Estimating Depth from Monocular Images as Classification Using Deep Fully Convolutional Residual Networks, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 28, No. 11, November 2018, pp. 3174 – 3182.

[19] D. Eigen, R. Fergus: Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture, Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, December 2015, pp. 2650 – 2658.

[20] J. L. Schönberger, E. Zheng, J.- M. Frahm, M. Pollefeys: Pixelwise View Selection for Unstructured Multi-View Stereo, Proceedings of the14th European Conference on Computer Vision, Amsterdam, Netherlands, October 2016, pp. 501 – 518.

[21] J. L. Schönberger, J.- M. Frahm: Structure-from-Motion Revisited, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, June 2016, pp. 4104 – 4113.

[22] N. Silberman, D. Hoiem, P. Kohli, R. Fergus: Indoor Segmentation and Support Inference from RGBD Images, Proceedings of the 12th European Conference on Computer Vision, Florence, Italy, October 2012, pp. 746 – 760.

[23] M. Liu, M. Salzmann, X. He: Discrete-Continuous Depth Estimation from a Single Image, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, June 2014, pp. 716 – 723.

[24] H.- S. Fang, G. Lu, X. Fang, J. Xie, Y.- W. Tai, C. Lu: Weakly and Semi Supervised Human Body Part Parsing via Pose-Guided Knowledge Transfer, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, June 2018, pp. 70 – 78

[25] L.- C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille: DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 40, No. 4, April 2018, pp. 834 – 848.

[26] L.- C. Chen, Y. Yang, J. Wang, W. Xu, A. L. Yuille: Attention to Scale: Scale-Aware Semantic Image Segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, June 2016, pp. 3640 – 3649.

[27] K. Simonyan, A. Zisserman: Very Deep Convolutional Networks for Large-Scale Image Recognition, Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, USA, May 2015, pp. 1 – 14.

[28] AgiSoft MetaShape Professional (Software), Version 1.5.1, 2016, Retrieved from: http://www.agisoft.com/downloads/installer/