ON THE PERFORMANCE OF DIBR METHODS WHEN USING DEPTH MAPS FROM STATE-OF-THE-ART STEREO MATCHING ALGORITHMS

Adriano Q. de Oliveira, Thiago L. T. da Silveira, Marcelo Walter and Cláudio R. Jung

Institute of Informatics, Federal University of Rio Grande do Sul, Brazil

ABSTRACT

In this paper we compare the quality of synthesized views produced by four DIBR methods when fed by depth maps estimated by five state-of-the-art stereo matching algorithms. Also, we compute the correlation between four popular metrics for ranking stereo matching algorithms and two metrics commonly used to evaluate synthesized views (PSNR and SSIM) plus one specific for DIBR. Among our findings, we highlight that (i) PSNR and SSIM have a weak correlation with common stereo matching metrics, (ii) using ground-truth depth does not lead necessarily to the best DIBR result; and (iii) estimated depth maps present artifacts that affect differently DIBR methods.

Index Terms— quality assessment, stereo matching, view synthesis, depth-image-based rendering (DIBR)

1. INTRODUCTION

Stereo matching (SM) is a well-studied problem, and it has been applied to several research-linked tasks such as robot navigation, surveillance and obstacle detection [1]. More recently, novel gadgets like Apple iPhone Xs and Samsung Galaxy S9+, which provide a portable easy-to-use built-in stereo vision camera setup, are bringing this technology to the end user. In this context, 3D photography – achievable by exploring the two cameras – is a promising way for recording and storing view-point changing still images and videos.

However, synthesizing novel views when the view-point is far from the original capture positions is still an open problem [2]. One particular class of view synthesis approaches is based on depth-image-based rendering (DIBR) [3], which uses as input a *single color image and its associated depth map* (obtainable via SM), and produces a novel synthesized view. To produce coherent novel views, DIBR methods must deal with occlusions/disocclusions, out-of-field areas (OOFAs), ghosts and cracks [4].

Several methods that address the DIBR problem [5, 6, 7, 8, 9, 10]. However, to the best of our knowledge, these methods assume that the depth (or disparity) maps are provided, i.e., they use *ground-truth* depth maps for both quantitative and qualitative assessment. Unfortunately, despite the increasing advances in both acquisition techniques and algorithms along the last years, the assumption of having highly accurate depth maps is still unrealistic for most practical applications [11]. Thus, the performance of each DIBR method in real scenarios, for which the disparity map must be estimated, is practically unknown.

The present study aims to evaluate the quality of the synthesized views produced by different DIBR approaches when fed with realistic disparity maps produced by SM approaches. It also aims to answer the following research question: "Are the stereo matching and view synthesis evaluation metrics correlated?" In other words, we want to know if, consistently, a "well-ranked" SM algorithm (according to a given SM metric) will provide better results in the DIBR context (according to a DIBR metric) than another "poorly ranked" method. Aiming to make our analysis as complete as possible, we select five SM algorithms [1, 12, 13, 14, 15] ranked according to four commonly used metrics [16]. On the other hand, four methods for DIBR [5, 7, 6, 8] are considered for comparison. We relate them based on a figure-of-merit that is specific for assessing DIBR-synthesized views [17].

The rest of this paper unfolds as follows. Section 2.1 revises a few closely related works. Then we expose the protocol for selecting SM and DIBR methods for our analysis, and briefly explain them in Sections 2.2 and 2.3, respectively. Supported by our hypotheses, the experimental setup is shown in Section 3. Section 4 discusses the obtained results. Finally, we conclude the paper in Section 5.

2. RELATED WORK

2.1. Quality Assessment Works

Here we discuss the closely related works that investigate the relationship between the problems of stereo matching and view synthesis. Lu and colleagues [18] found that the root mean square (RMS) error of estimated disparity maps may not correlate with the quality of interpolated views. These conclusions are drawn from experiments varying the SM algorithms but using only one *view interpolation* (VI) method. Taking into account the issues commonly found in VI pipelines, they propose a novel figure-of-merit for ranking SM algorithms. In practice, VI and DIBR methods do not suf-

This study was partially funded by the Coordenação de Aperfeiçoamento de Pessoal de Nvel Superior - Brasil (CAPES) - Finance Code 001 - and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

fer from exactly the same issues, so that their metric does not apply to this work. It is worth mentioning that VI methods require at least two color and two depth images, differing from the DIBR approach which uses only one color-plus-depth (V+D) image.

Similarly, Fuhr *et al.* [19] compared different SM algorithms facing VI as their target application. They consider different SM algorithms, chosen ad-hoc, but only one VI method was tested, as done in [18]. Furthermore, a single figure-of-merit is used for ranking SM algorithms, and the novel views are assessed by general-purpose image quality metrics. They concluded that the "number of bad pixels" in estimated disparity maps, which is a common metric for evaluating SM methods, is weakly correlated to the peak signal-to-noise ratio (PSNR) [20] and structural similarity index (SSIM) [21] measurements from the synthesized views.

To the best of our knowledge, no other studies are analyzing the impact of estimated depth maps on the results of view synthesis (VI or DIBR) methodologies.

2.2. Stereo Matching Algorithms

Stereo matching simulates the functioning of human eyes to obtain the distance from objects to two slightly shifted cameras by computing the disparities of corresponding pixels [1]. There is growing a number of SM algorithms, and selecting the one which performs best for a given task is not trivial [19]. Also, different quantitative error metrics capture different error types. In this work, we select four SM metrics used in the well-known Middlebury benchmark [16], namely `bad 2.0`, `avgerr`, `rms` and `a95`. They form a representative subset from all the metrics in Middlebury, favoring different error types in the matching.

Errors based on the percentage of "bad" pixels classify pixels into good and bad based on an acceptance threshold for the disparities differences. Bad pixels have the same penalty regardless of how far they are from the actual value. This is the case of the metric `bad 2.0`, where the threshold is set to 2.0. On the other hand, sums/means of absolute or squared errors increasingly penalize estimates far from the ground-truth, so that few bad pixels can corrupt the result. Metric `avgerr` captures the average absolute error, `a95` the 95th percentile error, and `rms` the RMS error, all of them in pixels. We sort the algorithms listed in the benchmark[1] in ascending order, in which their ranking score is given by the sum of the respective positions according to `bad 2.0`, `avgerr`, `rms` and `a95` for the test dense set and disregarding occlusion-removal masks. The smaller the score of the combined metric the better. Then, we select the five best performing algorithms with source code available according to the combined metric. They are briefly explained next.

Yin *et al.* [1] present a dictionary learning data-driven matching cost approach for comparing image patches. Instead of relying on color or texture information from the patches, their method computes the dissimilarity between learned sparse codes obtained from the input views. Those matching features are further incorporated in a semi-global cost aggregation and a post-processing step.

Zhang and others [14] propose a global SM algorithm that works on a 2D triangulation of the reference view, and generates a surface mesh containing depth information that is suitable for rendering virtual view-points (VVPs). Their two goal tasks, disparity estimation and view interpolation, are modeled as a two-layer Markov Random Field (MRF), being enforced each one by a separate layer. The authors claim that both problems are connected, and they indeed show compelling results. Despite that, the source code released by the authors does not implement the 2D triangulation, but instead *only* super-pixel segmentation via SLIC.

Taniai et al. [13] proposed a global optimization method also based on MRF with a continuous stereo approach. In their work, an initial randomized solution is optimized by MRF formulated by two terms: one that measures the photo-consistency between pixels, and another that aims to penalize local disparity discontinuities. For each patch of a grid-formatted image, local expansion moves are applied considering localization and spatial propagation in a graph-cut optimization scheme. This process is applied using a randomized search to infer continuous label space.

Mozerov and Weijer [12] explore the potential of cost filtering and energy minimization in their method. Their cost volume is defined by a linear combination of two per-pixel dissimilarities between left and right stereo images and their gradients. To generate the final disparity map, two distinct MRF models are used. A fully connected model is used in energy minimization of the cost filtering, and then a locally connected model is employed to exploit unary potentials. Finally, they perform a series of post-processings. The same authors also proposed another method, that they name OVOD [15]. The article associated to OVOD is in peer review process, with no preprint version, so that its definition is not accessible. Since OVOD meets the requirements of our ranking approach, we consider it in our analysis.

2.3. Depth-Image-Based Rendering Algorithms

The core idea of a DIBR pipeline is to project a real (reference) image to a desired VVP via 3D warping [6, 7]. In this process, however, different types of artifacts may appear in the synthesized view, such as cracks, ghosts, disocclusions, and OOFAs [4]. Cracks are related to rounding errors in the warping estimation process [5, 20]. Disocclusions are background areas occluded by foreground objects in the reference image that should be visible in the VVP [22]. Ghosts appear on depth discontinuities that are not sharp enough in the image domain [8]. Finally, OOFAs arise when the VVP exceeds the reference view bounds, creating regions without information on the edges of the synthesized view [4]. We consider

[1] Algorithms ranked in `http://vision.middlebury.edu/stereo/eval3` up to the paper submission.

four open-source DIBR algorithms that tackle those artifacts differently. Unlike in SM, there is no standardized protocol for comparing DIBR methods.

Solh and AlRegib [6] introduced a hierarchical hole filling (HHF) algorithm that can be coupled to any DIBR pipeline. The main idea of the authors is to process the warped image (synthesized view) in a coarse to fine approach, filling missing data (holes) with neighboring pixels via pseudo-zero canceling plus Gaussian filtering. They also propose a depth adaptive version of HHF that weights pixels according to how far they are from the camera. The code released by the authors does not implement the depth adaptive HHF version.

Unlike HHF that treats all warping artifacts simply as "holes", Ahn and Kim [5] make distinction between empty cracks and ghosts. In their method, although ghosts are identified they are marked as arbitrary holes. Furthermore, empty cracks are completed by median filtering. The authors rely on an extension of the well-known Criminisi's exemplar-based inpainting algorithm [23] for completing the remaining holes. The hole filling order is defined based on a confidence term that prioritizes background points, and patch matching search is performed in a limited region considering only background areas. The most similar patch in color sense provided that the depth in this region is "nearly flat" is chosen.

The selective hole-filling (SHF) method proposed in [8] identifies and corrects cracks and ghosts, and tackles larger holes by exploring depth in a patch-based inpainting scheme. More precisely, ghost points are moved to the opposite edge of the hole (foreground), and empty cracks are filled by a fast inpainting algorithm [24]. Afterwards, an extension of Criminisi's algorithm is used to complete the remaining holes. In this extension, the data term is replaced by a new depth term that aims to fill the holes prioritizing smaller disparities.

More recently, besides empty cracks and ghosts, Oliveira and colleagues [7] proposed to tackle also translucent cracks, OOFAs and disocclusions separately. Thus, the method is fully artifact-type aware and, for short, hereafter we will refer to it as ATA. ATA's preprocessing is performed in two steps. Initially, points classified as ghosts are reprojected to their correct position according to the estimated disparity value in the foreground. Then, translucent and empty cracks are detected and completed using the HHF algorithm. Unlike to [8, 5], disocclusions and the OOFAs are filled by two different extensions of the Criminisi's algorithm. For both hole types, the search is performed in a delimited region on the original image, but using dynamically adaptive patch sizes.

3. EXPERIMENTAL SETUP

Here, we evaluate the performance of all the selected DIBR methods when using depth estimated by all the considered SM algorithms (plus the ground-truth disparity, which is available in [16]). Parameters of the methods are set according to the papers or, if missing, according to the released source codes. Metrics for ranking SM algorithms were presented in Section 2.3. For assessing synthesized views, we use PSNR and SSIM since they are the standard figure-of-merit used in [6, 5, 8, 7], and also the context-specific morphological-wavelet PSNR (MW-PSNR) [17]. Since we need at least three views, two for estimating depth and a third one for assessing the synthesized view, we use the multi-view half-sized image sets from Middlebury 2006 [16].

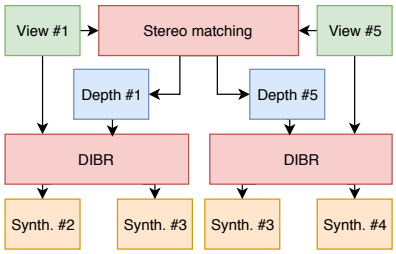

Fig. 1. Pipeline of our experimental setup.

Fig. 1 depicts the pipeline of our experimental setup. We feed *Views 1* and *5*, which have ground-truth for the disparity map, to a given SM algorithm and produce two disparity maps, one for each reference view (*1* and *5*). Then, we recover depth from the camera intrinsics, given in [16], and the disparities. Together with the color image, we submit the estimated depth maps to a selected DIBR technique. We synthesize novel views in the same camera positions of real *Views 2* and *3* based on the V+D information of *View 1*, and *Views 3* and *4* from the V+D data of *View 5*. This way we can assess the quality of both depth maps and synthesized views against the respective ground-truths.

4. RESULTS AND DISCUSSION

Based on the pipeline explained in Section 3, we are able to compare the results varying (i) SM and (ii) DIBR techniques, being assessed via figures-of-merit such as (iii) `bad 2.0`, `avgerr`, `rms` and `a95`, and (iv) PSNR, SSIM and MW-PSNR. Note that relating those variables form a four-dimensional hyper-cube. Moreover, there are two additional dimensions: the image sets from [16], and the two depth maps and four synthesized views per image set.

Table 1 presents average PSNR, SSIM and MW-PSNR results (for the $21 \times 4 = 84$ views of the dataset) for all the DIBR methods and SM algorithms besides the ground-truth depth map (column GT). The best results for each metric are highlighted in boldface. Note that using GT disparity maps *does not* lead to the best synthesized view. In fact, using methods like [1, 15, 13] for estimating depth tends to produce better PSNR, SSIM and MW-PSNR measurements than if the ground-truth is used. Similar findings were also reported in [18]. Contrarily to what occurs with the selected SM algorithms, ground-truth maps have no disparity values in regions classified as unknown [16], i.e., regions without correspondences in both left and right views. In practice, DIBR methods have much more pixels to estimate when using the

Table 1. Average PSNR, SSIM, and MW-PSNR per row. Columns present the SM algorithms (and ground-truth depth map), whereas the triple-rows delimit the DIBR methods.

	[12]	[1]	[14]	[15]	[13]	GT
	24.334	31.960	16.883	31.143	30.918	29.273
[6]	0.7276	0.9532	0.5331	0.9527	**0.9536**	0.9472
	26.258	**32.926**	20.207	32.593	32.025	31.475
	24.548	31.335	18.254	31.024	30.128	28.713
[5]	0.7174	0.9393	0.5258	0.9427	0.9402	0.9290
	26.225	30.892	20.363	31.419	30.478	30.075
	24.685	32.196	18.230	31.793	31.811	30.066
[8]	0.7229	0.9450	0.5314	0.9496	0.9499	0.9422
	25.881	28.854	20.080	30.050	28.951	29.649
	24.859	**32.626**	18.292	32.052	32.041	31.721
[7]	0.7236	0.9512	0.5368	0.9517	0.9531	0.9522
	26.435	32.848	20.432	32.717	32.330	32.609

ground-truth depth map.

One may also note that PSNR, SSIM and MW-PSNR suggest three different rankings for DIBR methods when ground-truth depth map is used. Moreover, if instead we estimated depth, then we may end up with another ranking. More precisely, the rankings based on PSNR, SSIM and MW-PSNR are inconsistent w.r.t. that based on the depth ground-truth for methods [14], [15, 13, 12] and [1, 14], respectively.

The relative ranking order between the considered SM algorithms according to `bad 2.0`, `avgerr`, `rms` and the combined score of all analyzed metrics is the same: LocalExp [13], OVOD [15], MeshStereo [14], DDL [1] and TSGO [12]. The ranking order based on metric `a95` exchanges the first and second best-performing algorithms, and also the two last ones. If using the view synthesis metrics for ranking, we end up with different orders, as can be seen in Table 1. Specifically, the method in [14] performed worst, probably because the released code is a rough approximation for the actual published method.

We further investigate the relationship between SM and DIBR metrics. Table 2 presents the strongest Spearman correlation [25] and the corresponding SM and DIBR techniques, for all the combinations of SM and DIBR metrics. One may note that our results suggest that metrics `bad 2.0` and MW-PSNR have a fairly strong negative relationship. This analysis also indicates that *it is not expected* to have necessarily higher SSIM and PSNR values for synthesized views when we choose SM methods that minimize the error metrics `bad 2.0`, `avgerr`, `rms` and `a95`. Our findings agree with those shown in [19] for the view interpolation scenario.

DIBR artifact types reported in the literature are related to ground-truth depth maps. SM-based depth maps may differ, leading DIBR techniques to present contrasting results. By visual inspection we could note that cracks, ghosts, and OOFAs presented the same patterns reported in literature, although the first two tend to appear more intensely. However, disocclusion regions are contaminated with over-segmented depth layers due to inaccurate depth estimation, producing

Table 2. Correlation analysis for SM and view synthesis metrics. References within the cells indicate the methods for which the maximum correlation was achieved.

	`bad 2.0`	`avgerr`	`rms`	`a95`
PSNR	$-0.39^{[15,\,8]}$	$-0.37^{[15,\,8]}$	$-0.34^{[14,\,8]}$	$-0.40^{[15,\,8]}$
SSIM	$-0.40^{[14,\,8]}$	$-0.47^{[14,\,5]}$	$-0.44^{[14,\,5]}$	$-0.33^{[14,\,5]}$
MW-PSNR	$-0.80^{[13,\,8]}$	$-0.74^{[13,\,8]}$	$-0.68^{[13,\,8]}$	$-0.65^{[13,\,8]}$

a distortion effect especially in foreground objects and their edges. The effect of over-segmentation in ghost areas tends to be small for techniques based on successive averaging of local information, such as HHF [6]. Differently, for patch-based techniques [5, 8, 7], this problem may produce many incoherent artifacts. These techniques perform the reconstruction process using a patch as a model (selected from the target hole edge), which is compared to valid information in the synthetic view, and the most similar is copied to the empty region, iteratively. Also, the patch-based techniques classify disocclusion edges based on the warped depth map in order to use models composed of background information. Nevertheless, the classification itself may not be satisfactory as the depth is not consistent with the texture, as illustrated in Fig. 2. More results will be available at the author's webpage: `http://www.inf.ufrgs.br/~mwalter/dibrxsm/`.

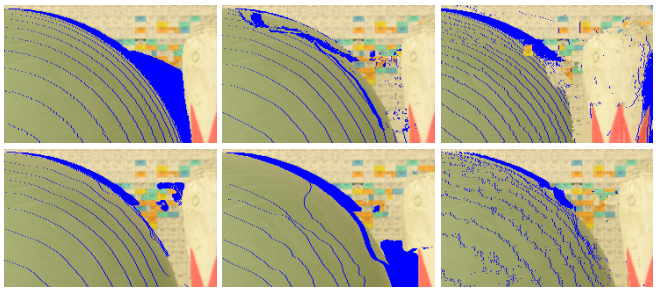

Fig. 2. Warped view of `Bowling1` using ground-truth and estimated depth-maps using [13], [14], [15], [12] and [1], from the top-left to bottom-right.

5. CONCLUSIONS

We presented a comparative study focusing on the quality of synthesized views produced by different DIBR techniques, coupled to depth maps estimated via different state-of-the-art SM algorithms. We have experimentally shown that: (i) DIBR methods can generate better results if using SM-based depth maps, instead of the ground-truth; (ii) DIBR techniques are ranked differently when fed by depth maps generated with SM algorithms or ground-truth depth; (iii) SM methods that minimize SM error measures do not necessarily result in better synthesized views according to SSIM and PSNR; (iv) MW-PSNR has a strong negative correlation to SM metrics, and may be more useful for assessing DIBR methods than PSNR and SSIM; (v) and SM-based depth maps contain errors that mislead DIBR techniques, indicating that they may not be prepared for real scenario applications.

6. REFERENCES

[1] J. Yin, H. Zhu, D. Yuan, and T. Xue, "Sparse representation over discriminative dictionary for stereo matching," *Pattern Recogn.*, vol. 71, pp. 278 – 289, 2017.

[2] J. Gautier, O. Le Meur, and C. Guillemot, "Depth-based image completion for view synthesis," in *3DTV Conference*, 2011, pp. 1–4.

[3] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," in *Stereoscopic Displays and Virtual Reality Systems XI*, 2004, vol. 5291, pp. 93–104.

[4] S. M. Muddala, M. Sjöström, and R. Olsson, "Virtual view synthesis using layered depth image generation and depth-based inpainting for filling disocclusions and translucent disocclusions," *J. Vis. Commun. Image Represent.*, vol. 38, pp. 351–366, 2016.

[5] I. Ahn and C. Kim, "A novel depth-based virtual view synthesis method for free viewpoint video," *IEEE Trans. Broadcast*, vol. 59, no. 4, pp. 614–626, 2013.

[6] M. Solh and G. AlRegib, "Hierarchical hole-filling for depth-based view synthesis in FTV and 3D video," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 5, pp. 495–504, 2012.

[7] A. Q. de Oliveira, M. Walter, and C. R. Jung, "An artifact-type aware DIBR method for view synthesis," *IEEE Signal Process. Lett.*, vol. 25, no. 11, pp. 1705–1709, 2018.

[8] A. Q. de Oliveira, G. Fickel, M. Walter, and C. Jung, "Selective hole-filling for depth-image based rendering," in *IEEE ICASSP*, 2015, pp. 1186–1190.

[9] G. Luo and Y. Zhu, "Foreground Removal Approach for Hole Filling in 3D Video and FVV Synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 10, pp. 2118–2131, 2017.

[10] D. M. M. Rahaman and M. Paul, "Virtual View Synthesis for Free Viewpoint Video and Multiview Video Compression using Gaussian Mixture Modelling," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1190–1201, 2018.

[11] M. Poggi and S. Mattoccia, "Learning to predict stereo reliability enforcing local consistency of confidence maps," in *IEEE CVPR*, 2017, pp. 4541–4550.

[12] M. G. Mozerov and J. van de Weijer, "Accurate stereo matching by two-step energy minimization," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 1153–1163, 2015.

[13] T. Taniai, Y. Matsushita, Y. Sato, and T. Naemura, "Continuous 3D label stereo matching using local expansion moves," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2018.

[14] C. Zhang, Z. Li, Y. Cheng, R. Cai, H. Chao, and Y. Rui, "Meshstereo: A global stereo model with mesh alignment regularization for view interpolation," in *IEEE ICCV*, 2015, pp. 2057–2065.

[15] M. G. Mozerov and J. van de Weijer, "One-view-occlusion detection for stereo matching with a fully connected CRF model," `http://datasets.cvc.uab.es/OVOD/`, 2018.

[16] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *IEEE CVPR*, 2007, pp. 1–8.

[17] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "Multi-Scale Synthesized View Assessment Based on Morphological Pyramids," *J. Electr. Eng.*, vol. 67, no. 1, pp. 3–11, 2016.

[18] J. Lu, Q. Yang, and G. Lafruit, "Interpolation error as a quality metric for stereo: Robust, or not?," in *IEEE ICASSP*, 2009, number D, pp. 977–980.

[19] G. Fuhr, G. P. Fickel, L. P. Dal'Aqua, C. R. Jung, T. Malzbender, and R. Samadani, "An evaluation of stereo matching methods for view interpolation," in *IEEE ICIP*, 2013, pp. 403–407.

[20] Y. Mori, N. Fukushima, T. Fujii, and M. Tanimoto, "View generation with 3d warping using depth information for ftv," in *3DTV Conference*, 2008, pp. 229–232.

[21] 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, 2004.

[22] G. Luo, Y. Zhu, Z. Li, and L. Zhang, "A hole filling approach based on background reconstruction for view synthesis in 3D video," in *CVPR*, 2016, pp. 1781–1789.

[23] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, 2004.

[24] M. M. Oliveira, B. Bowen, R. McKenna, and Y. Chang, "Fast digital image inpainting," in *VIIP*, 2001, pp. 261–266.

[25] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.