2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Cost Volume Refinement for Depth Prediction

João L. Cardoso
Institute of Visual Computing and Human-Centered Technology
TU Wien
Vienna, Austria
jaliborc@cg.tuwien.ac.at

Nuno Gonçalves
Institute of Systems and Robotics
University of Coimbra
Portuguese Mint and Official Printing Office
Lisbon
nunogon@deec.uc.pt

Michael Wimmer
Institute of Visual Computing and Human-Centered Technology
TU Wien
Vienna, Austria
wimmer@cg.tuwien.ac.at

***Abstract*—Light-field cameras are becoming more popular in the consumer market. Their data redundancy allows, in theory, to accurately refocus images after acquisition and to predict the depth of each point visible from the camera. Combined, these two features allow for the generation of full-focus images, which is impossible in traditional cameras.**

Multiple methods for depth prediction from light fields (or stereo) have been proposed over the years. A large subset of these methods relies on cost-volume estimates – 3D objects where each layer represents a heuristic of whether each point in the image is at a certain distance from the camera. Generally, this volume is used to regress a depth map, which is then refined for better results. In this paper, we argue that refining the cost volumes is superior to refining the depth maps in order to further increase the accuracy of depth predictions. We propose a set of cost-volume refinement algorithms and show their effectiveness.

***Index Terms*—Depth Reconstruction, Light-Fields, Stereo, Optimization, Cost-Volumes**

I. Introduction

Light field cameras, also called plenoptic cameras, are cameras that capture both the light direction and intensity emanating from a scene simultaneously. Typically, this is implemented as an array of micro-lenses placed in front of a conventional image sensor [3], [4]. They first became popular among professional photographers, due to their ability to precisely refocus images after acquisition [5], [6].

The data redundancy from the multiple micro-lenses also allows, in theory, to predict the depth of the scene from the camera. Yet, while image refocusing in plenoptic cameras is widely understood, depth reconstruction is still an active field of research, with multiple competing methods. When compared to multi-camera systems, the major limitation of light field cameras for depth reconstruction is that all their lenses are extremely close together. This results in very narrow baselines [7], [8]. Thus, typical multi-camera depth-reconstruction techniques are not appropriate to use with light-field imagery. On the other hand, light-field cameras are generally cheaper than multi-camera setups, more portable, and need no synchronization between different cameras. This has sparked an interest in depth-reconstruction methods specific for light fields.

In this work, we focus on a class of methods that rely on cost-volume estimates. We now explain the basic concepts surrounding cost volumes and how they are generally used. Let

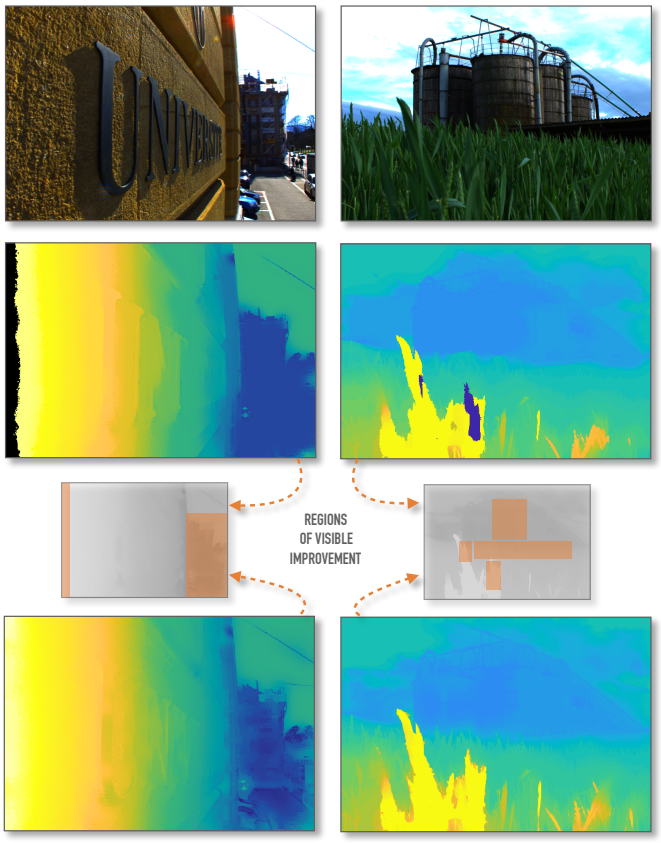


Fig. 1: Comparison of our results (bottom) with the ones obtained using Jeon *et al.*'s [1] proposed pipeline (centre) on two different light-field images from the dataset by Rebarek and Ebrahimi [2]. Notice that our proposed refinements are less prone to artifacts, and better preserve details of far away objects (the grass and metallic cylinders on the top, the street on the bottom).

$C(\mathbf{u}, z)$ be some cost volume, a three-dimensional function parameterized by the image coordinates u and depth z that, when minimized along the depth axis, should result in an accurate prediction D_C of ground-truth depth D:

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

$$D(\mathbf{u}) \simeq D_C(\mathbf{u}) = \underset{z}{\operatorname{argmin}}\, C(\mathbf{u}, z) \tag{1}$$

The pipelines used by cost-volume based depth prediction methods can be generalized to 5 stages: first, a cost volume is estimated from some cue(s) in the light field. Second, the cost volume might be refined to improve further predictions. Third, a depth map is estimated from the cost volume. Equation 1 is the simplest possible depth estimation method, but more complex methods have been proposed. Fourth, the depth method might be refined using refinement methods for range images. Finally, camera properties and calibration are taken into account to compute a depth map from the depth. We illustrate all stages in Figure 2 and how some existing work fits into them.

To the best of our knowledge, the majority of existing work focuses on cost estimation and depth estimation from the cost, which are arguably the most important steps. To compensate for limitations of methods used in those two stages, authors also often makes use of existing range-image refinement techniques. However, we believe this is a sub-optimal approach, as during this stage most of the cost-volume information (and thus, light-field redundancy) is no longer available. Instead, we propose focusing refinement efforts in the cost-refinement stage. Our main contributions are:

1) A modular framework for cost-volume refinement, which can be applied for depth reconstruction on light-fields, regular and multi-view stereo imagery.
2) A floating-point method for artifact removal on cost volumes based on classification methods robust to smooth surfaces and object complexity.
3) A fast local smoothing method for noise and discontinuity reduction on cost volumes robust to sharp depth changes.
4) A method for combining cost-volume based depth prediction with other prediction methods before regression.
5) Extensive testing of the importance of cost-volume refinement and of the efficacy of our methods on multiple previously proposed cost cues.

II. Related Work

Depth estimation from light fields has been an active research topic over the past few years. We focus primarily on methods that rely on cost volumes to regress depth, for which a large body of recent research already exists. However, it is important to point out that recent deep learning based methods have also shown promise. For example, Shin *et al*. [9] proposed a fully convolutional network capable of estimating depth from epipolar images. Zhou *et al*. [10] proposed three unsupervised loss functions, which remove the need for large amounts of ground truth data for training.

The most commonly used cues for cost volume estimation are defocus, correspondence and epipolar plane analysis. Defocus measures the optimal local sharpness for a given focus distance, which can be estimated after refocusing the image at different depths [11]. Tao *et al*.2015 [12] proposed adaptive defocus response, an extended cue more robust to occlusion. Williem *et al*.2018 [13] took the method one step further and proposed constrained adaptive defocus, a cue invariant to noise and occlusion.

Correspondence refers to the process of finding matching points on different sub-aperture images that represent the same point in the scene. However, light-field images generally have very narrow baselines, which cause stereo correspondence matching to obtain sub-par results due to sub-pixel shift [14]. Thus, correspondence generally refers to angular patch-based estimation methods, even though standard multi-view stereo data cost, calculated from the sum of absolute differences, is also used [12]. Jeon *et al*. [1] used the phase-shift theorem to estimate sub-pixel shift in the image frequency domain. Tao *et al*.2013 [11] estimated correspondence as the variance in the angular patch of refocused images. As they did for the defocus cue, Williem *et al*.2018 [13] proposed constrained angular entropy, a cue invariant to noise and occlusion.

Related to the concept of angular estimation is epipolar plane image analysis. It refers to slicing the light field along the epipolar planes, and taking advantage of properties of the resulting images to estimate properties [15]. In particular, epipolar images tend to form diagonal lines whose angles are linearly related to depth [16]. However, these lines only allow for sparse estimations, making them sub-optimal for dense cost volume generation.

Yet our focus is not how the cost volumes are generated, but how they are processed and how depth is regressed from them. Depth estimations from previous work often exhibit sharp discontinuities, which are particularly problematic in flat and smooth surfaces. This occurs due to a strict trade-off between computational efficiency and cost precision, as both are tied to the number of layers of the volume. For example, classification methods can solve ambiguities in the cost volume, in particular in out-of-focus background regions, and thus remove unwanted artifacts. However, they can also further exacerbate discontinuity artifacts by reducing the number of possible depths to a limited set of classes, or create new artifacts of their own due to mislabeling. Both Jeon *et al*. [1], [17] and Williem *et al*. [13] suffer from this issue, as they perform multi-label optimization, using graph cuts [18], to propagate SIFT [19] feature matches.

To compensate limitations of the regression methods, previous work often performs refinement operations on the disparity or depth maps obtained from regression. After their multi-label regression, both Jeon *et al*. [1], [17] and Williem *et al*. [13] perform median weight transfer, followed by an iterative spatial-depth super-resolution method first proposed by Yang *et al*.2007 [20].

Different cues can also be integrated. For example, Jeon *et al*. [17] uses four different cues, which are used to generate a total of 16 cost volumes, to compensate for each other's shortcomings. Defocus and correspondence are the most often combined [21]–[23]: Tao *et al*. [11] first combined these two cues using Markov random field propagation with the Peak Ratio, introduced by Hirschmüller *et al*. [24], as the confidence

measure. Later, Tao *et al*. [12] improved it by also using a shading constraint as the regularization term.

However, depth refinement methods are unable to take full advantage of the light-field properties, as they are reduced to work in 2D color and depth space. Due to the lack of information, we found they are prone to creating artifacts or misreading the shape of the scene.

III. Minimal Pipeline

To test our proposed cost-volume refinements, we present a simplified pipeline, which abstains from complex depth regression and depth refinement techniques. This is done for two reasons: first, one of the advantages of cost refinement is the reduced necessity of such techniques, which we want to show. Second, some existing regression methods, such as graph cuts label propagation [1] [13], cause a great loss of detail and a portion of our improvements could be lost.

All of our results are computed using this minimal pipeline. For comparison, results of previous work are always computed using their respective original pipelines, shown in Figure 2.

As explained in Section I and shown in Figure 2, all of these pipelines can be generalized to 5 stages, and ours is no exception. First, for cost-volume generation, we use cues from existing work. Then, for each volume, our refinement methods are applied in succession: obvious artifacts are removed using our classification-based global artifact removal, described in Section IV-B. Noise and unwanted sharp discontinuities are vastly reduced using our iterative local smoothness refinement, described in Section IV-C. Optionally, depth predictions from non cost-volume based methods can also be combined in this stage, as described in Section IV-A.

After our proposed refinements have been applied, we make use of a classical solution to estimate depth from the cost. The theoretical depth regression, described in Equation 1, assumes that cost is a continuous function. However, cost volumes are computed and stored in discrete steps. Reducing this step ξ increases depth precision but at the cost of computational performance, and precision is required to effectively use cost refinement. Thus, we use parabolic interpolation [25] [26], which takes into account information from the immediate neighbors of the minimum step $D_C(\mathbf{u})$:

$$D(\mathbf{u}) \simeq \mathcal{D}_C(\mathbf{u}) = D_C(\mathbf{u}) - \left(1 + 2 \cdot \frac{C(\mathbf{u},\, D_C(\mathbf{u}) - \xi) - C(\mathbf{u},\, D_C(\mathbf{u}))}{C(\mathbf{u},\, D_C(\mathbf{u}) + \xi) - C(\mathbf{u},\, D_C(\mathbf{u}) - \xi)}\right)^{-1} \tag{2}$$

This simple solution highly increases depth precision without increasing computational costs.

IV. Refinement Algorithms

Let $k \in [0, n_r[$ be the number of cost-volume refinements that have been performed by a pipeline, where n_r is the total number of refinement algorithms being used. We generalize modular refinement as a function R_k that takes as input the current cost volume C_k and outputs a refined volume C_{k+1}:

$$C_{k+1}(\mathbf{u}, z) = R_k(C_k,\, \mathbf{u}, z,\, \lambda_k) \tag{3}$$

In this work, we present three different modular refinement algorithms, all of which follow the definition of Equation 3 and can be used interchangeably in any order. The strength of refinement R_k can be controlled with hyper-parameter $\lambda_k \in \mathbb{R}$.

A. Independent Predictor Combination

We first describe our simplest refinement. Our goal is to define a generic method that can make use of any independent depth prediction to inform our own. Independent predictions can be obtained from non cost-based light-field methods, from existing methods for monocular or stereo imagery, or from any domain specific knowledge.

To do so, we increase the original cost according to the difference between depth and the independent method's prediction. There should be no increase when in agreement, but cost should increase as the two diverge. Additionally, the increased cost should have a known maximum $\in \mathbb{R}$, so that the refinement impact can be controlled. Given these properties, we define the increased cost $G(t) \in [0, 1]$ as normalized inverted Gaussian distributions centered around the independent predictions:

$$G(t) = 1 - e^{-\frac{t^2}{2\sigma^2}} \tag{4}$$

where σ is a chosen deviation. For each image point u for which the independent method has a prediction P_u, the refinement operation becomes:

$$C_{k+1}(\mathbf{u}, z) = C_k(\mathbf{u}, z) + \lambda_k G(P_{\mathbf{u}} - z) \tag{5}$$

As an example, we explore the case of facial reconstruction. We make use of the trained neural network presented by Sela *et al*.2017 [27] for facial reconstruction from color images as our domain-specific prior knowledge. We estimate prior depth P from the neural network depth prediction and, for each pixel that we have a prior $P_{\mathbf{u}}$ for, we apply the refinement. The result can be seen in Figure 3.

B. Classification Artifact Removal

We propose a variation of the previous refinement for the purposes of artifact removal. In particular, as described in Section II, we found that some multi-label classification methods (where discrete steps in depth correspond to labels) are robust to artifacts, but tend to reduce accuracy due to lack of precision or miss-assignment between close labels. To take advantage of the artifact detection while maintaining accuracy, we define a increased cost different from Section IV-A.

As before, we increase cost according to the difference between depth and the multi-label classification. However, it should not be bound to a known maximum and should instead quickly rise with divergence. As such, we define the increased cost as a polynomial of degree m, where m is an even number. Additionally, artifacts are more likely the higher the difference between predicted parabolic depth $\mathcal{D}$ and

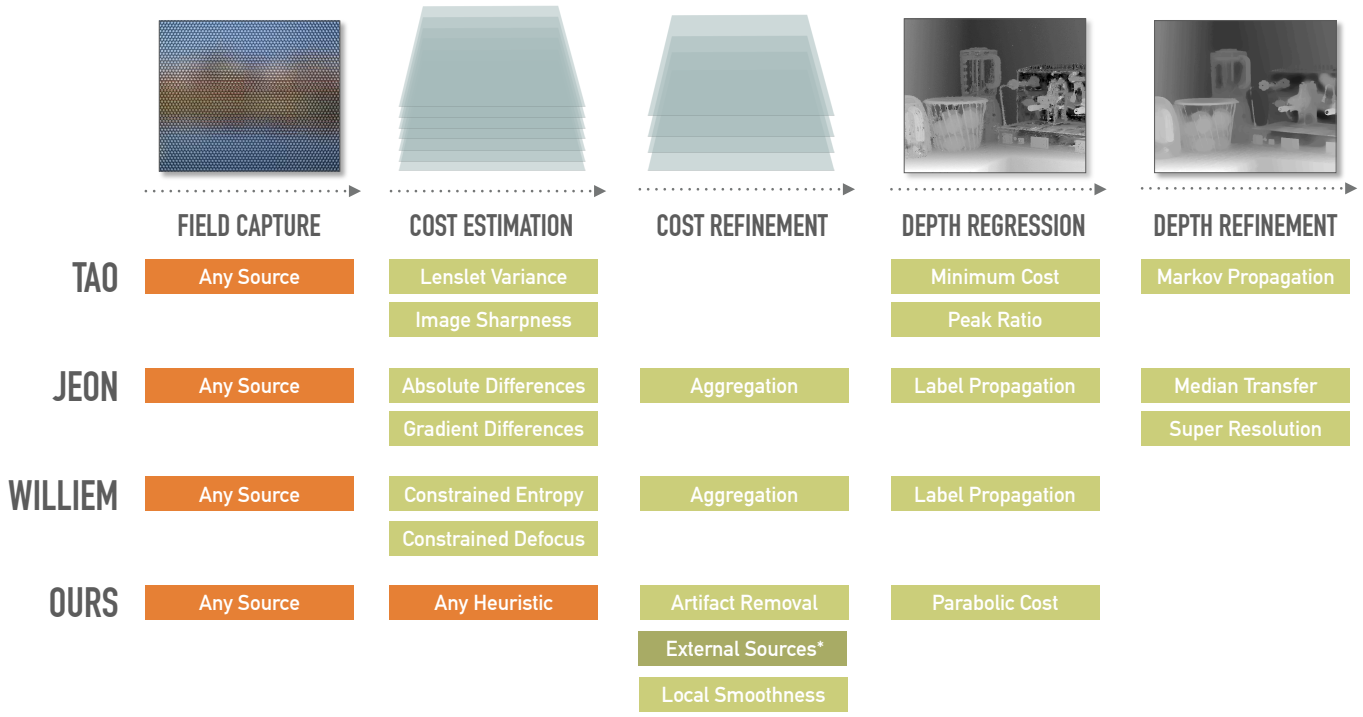


Fig. 2: An overview comparison of our minimal pipeline with Tao *et al.* [11], Jeon *et al.* [1] and Williem *et al.* [13]. Images are only illustrative. Note that, due to its specificity, the combination of external predictions or domain-specific knowledge, described in Section IV-A, is not used to generate the majority of the results in this paper nor when comparing accuracy in Section V.

multi-label classification L is. Thus, we scale the increased cost according to the difference between these two predictions computed from the current cost volume C_k.

Our refinement thus becomes:

$$C_{k+1}(\mathbf{u}, z) = C_k(\mathbf{u}, z) + \lambda_k |L_\mathbf{u} - \mathcal{D}(C, \mathbf{u})| \cdot (L_\mathbf{u} - z)^m \quad (6)$$

As an example, we use the graph-cuts implementation of Jeon *et al.* [1] for propagation of SIFT feature matches to estimate each label L_u at pixel u with $m = 2$. A direct comparison of our refinement to the original algorithm can be seen in Figure 4.

C. Iterative Local Smoothness

Two common issues with depth predictions from cost volumes are noise and local artifacts. We vastly reduce these by looking at the neighborhood $\mathcal{I}_u$ of each image point u. For each neighbor $v \in \mathcal{I}_u$, we create an added cost based on the difference between the depth predictions at v and u. To weight the importance of each neighbor, we estimate point confidence using the peak ratio coefficient W (as proposed by Hirschmüller *et al.* [24]), which produces lower values when multiple local minima are similar:

$$W_C(\mathbf{u}) = \frac{C(\mathbf{u}, \mathcal{D}_C(\mathbf{u}))}{C(\mathbf{u}, \operatorname{argmin}_{z \neq \mathcal{D}_C(\mathbf{u})} C(\mathbf{u}, z))} \quad (7)$$

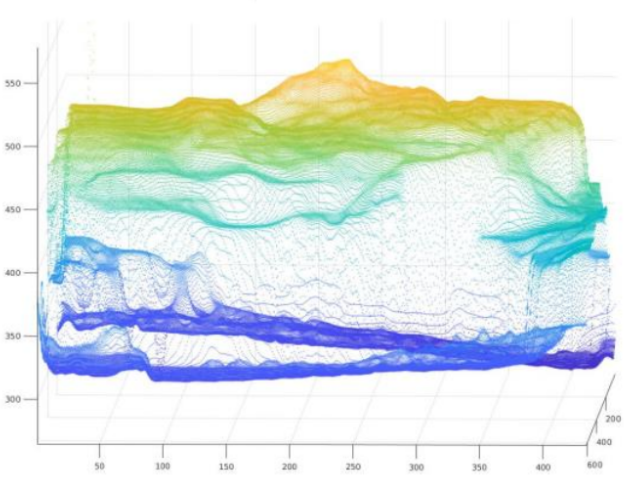
Just as for the refinement in Sections IV-A, we want agreeing predictions to have no additional cost, but to increase cost as local differences raise up to a chosen maximum. As such, we use the same normalized inverted Gaussian distribution G. However, we want to define the increased cost as a function of depth prediction at the neighbors, which is in turn dependent on the cost increase. To deal with the conundrum, we solve the problem by estimating new cost volumes iteratively. Let $j \in [0, n_i[$ be the current iteration number. We define new temporary volumes S as:

$$\begin{cases} S_0 = C_k \\ S_{j+1}(\mathbf{u}, z) = C_k(\mathbf{u}, z) + \\ \qquad \lambda_k \sum_{v \in \mathcal{I}_\mathbf{u}} G(\mathcal{D}_{S^j}(v) - z) \cdot W_{S^j}(v) \end{cases} \quad (8)$$

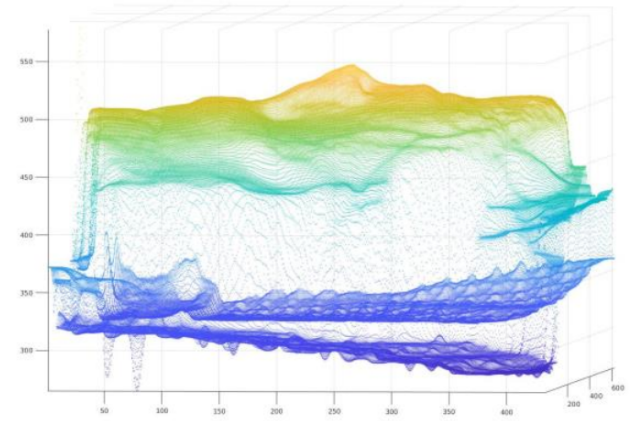
where $\mathcal{D}_{S^j}$ is the parabolic depth prediction and W_{S^j} the Peak Ratio coefficient computed from temporary cost volume S_j, according to Equations 2 and 7 respectively. n_i is the total number of iterations to be performed, which can be either statically or dynamically controlled by the pipeline. We make use of the Peak Ratio because different neighbors might have more or less reliable cost predictions than others, and thus should be weighted differently.

Having the last iteration been performed, we define our refinement as:

$$C_{k+1}(\mathbf{u}, z) = S_{n_i}(\mathbf{u}, z) \quad (9)$$

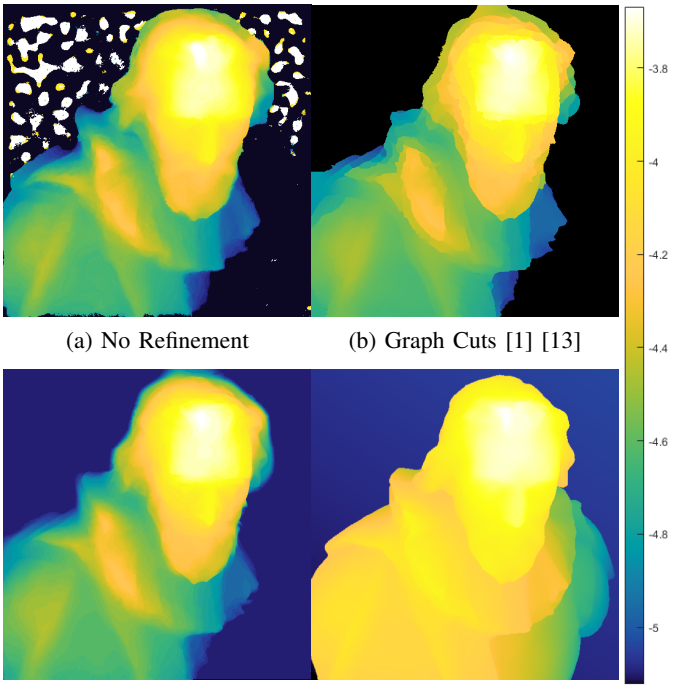


(a) Not Refined



(b) Refined

Fig. 3: Depth reconstruction from a light-field portrait picture with and without refinement from Section IV-A. We used our refinement to combine information from the facial neural network proposed by Sela *et al.*2017 [27].



(a) No Refinement (b) Graph Cuts [1] [13]

(c) Artifact Removal (ours) (d) Ground Truth

Fig. 4: Reconstructions of a bust on an intentionally poorly estimated cost volume, which overestimates depth and is unable to predict depth of far away objects, resulting in multiple visible artifacts. Graph cuts prediction (b) removes the background artifacts, but at the cost of depth accuracy of the bust itself. The artifact removal refinement (c) from Section IV-B outperforms previous methods, being able to both remove artifacts and reduce depth overestimation without decreasing accuracy.

In our implementation, we set a static maximum number of iterations, but we also monitor the ratio of change between depth maps $\mathcal{D}_{Sj}$ and $\mathcal{D}_{Sj+1}$. Once the ratio is below a predefined threshold, we stop iterating. We found that our implementation never requires more than 2 iterations before converging.

V. Results

To test the effectiveness of our refinement algorithms, we look at three very different cost-volume generation cues proposed by three different authors: Tao *et al.* [11] lenslet variance (LV), Jeon *et al.* [1], [17] sum of absolute differences computed using sub-pixel phase-shift (SAD) and Williem *et al.* [13] constrained angular entropy (CAE). We compare the depth maps regressed from these volumes with our minimal pipeline (as described in Section III), which includes cost volume refinement, to the ones regressed using the pipelines publicly provided by their authors (see Figure 2).

To do the comparisons, we use the synthetic dataset by Honauer [28], which contains 30 pairs of light-field images and corresponding ground-truth depth maps of different scenes. We also present a visual comparison using the dataset by Rebarek and Ebrahimi [2], as shown in Figures 1 and 6. For each scene, we generate cost volumes according to the three mentioned cues. Then, for each volume, we regress depth maps using our minimal pipeline and the originally corresponding one. This results in a total of 6 different depth maps per scene.

Note that both Tao *et al.* [11] and Jeon *et al.* [1], [17] combine multiple cues using weighted sums of different cost volumes in their works. However, we are not proposing an end-to-end depth prediction method, but a set of operations that, given an arbitrary cost volume, are able to generate a better and more consistent volume. Thus, we analyze the performance of refinement on different cues individually.

Additionally, the absolute error between predictions and ground truth of a specific example are not relevant, as they are largely constrained by the quality of the input cost volume. Instead, we look at how this error changes with the introduc-

Pipeline	MSE	SSI
LV (Tao *et al.* [11])	2.1672%	9.6871
Refined Lenslet Variance	1.5297%	10.8989
SAD (Jeon *et al.* [1])	1.2829%	11.3914
Refined Sum of Absolute Differences	0.7165%	11.6619
CAE (Williem *et al.* [13])	3.2723%	8.6063
Refined Constrained Angular Entropy	2.1083%	9.9078

TABLE I: Average of mean squared error (MSE) and structural similarity index (SSI) for each tested pipeline when predicting on the synthetic dataset by Honauer [28]. Lower MSE and higher SSI are better [28].

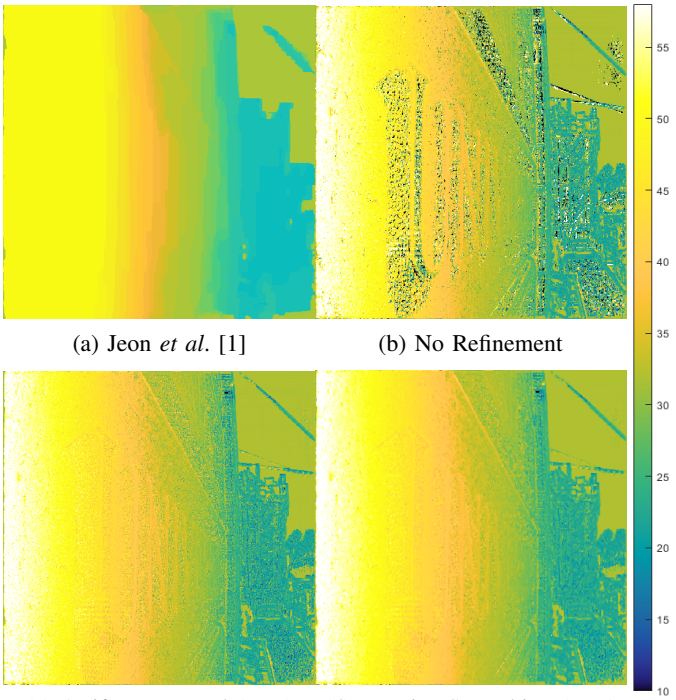
tion of refinements. Thus, Figures 5 and 7 display the change of error when our refinements (with our minimal pipeline) are used, color coded in green for reduced and red for increased error. Color is normalized to the highest change.

We do not tweak the configurable variables λ_k and σ from Equations 3 and 4 for each scene and use the same for all tests. We also do not perform any additional operations, such as vignetting and distortion estimation and correction, as these should affect all 6 cases equally, and operations before cost-volume generation are out of scope of this work.

a) Statistical Analysis: We calculate the mean squared error and the structural similarity index between the ground-truth maps and the regressed ones. Table I shows the average of these metrics for each of the 6 combinations. Our minimal pipeline outperforms the original ones in all cases, presenting a lower average error and higher similarity, even though it is not performing complex depth regression or depth refinement. We also found that our refinements are the more effective the worse the cost prediction is. For example, the differences are more visible in real photographs than synthetic (perfect) images, or in intentionally poor reconstructions. As such, the real error and similarity differences might be higher than suggested by our synthetic dataset.

b) Detail Preservation: As displayed in Figure 6, alternative proposed methods often oversimplify depth predictions.

Fig. 5: Comparison of the absolute per pixel error computed between ground truth Figure 4.d and predictions 4.b and 4.c. Green represents a lower error from our method, red an higher, with color being normalized to highest change. Section V elaborates on the reasoning behind the metric.



(a) Jeon *et al.* [1] (b) No Refinement

(c) Artifact Removal (ours) (d) Iterative Smoothing (ours)

Fig. 6: Depth regression before any image-based refinement is applied on a real-life light-field image from a dataset [2]. Previous work often forfeits detail and accuracy (a) to reduce artifacts and noise. Our use of cost volume refinement methods (c,d) from Sections IV-B and IV-C solve these issues while preserving detail. Depth was predicted using parabolic interpolation (b,c,d).

Most detail can be lost and only regained through depth-map refinement, which does not take advantage of the light-field properties over traditional images. Our algorithm is able to vastly reduce unwanted artifacts and noise, while preserving the details present in the scene by performing operations at a cost volume level. As such, the shape of objects in the final results more closely resembles the ground truth than previous methods, as shown in Figure 7.

c) Limitations: While functional, our smoothing refinement is still not able to remove all noise without removing details. As shown in Figure 6, the refined result still exhibits some noise. We also do not have a solution for optical effects such as flares, which can mislead predictions locally, as also shown in Figure 6. However, we are not aware of any existing depth refinement method to deal with this issue, and previous work frequently suffers from the same issue.

VI. Conclusion

We presented a novel approach for cost volume optimization for depth prediction, which relies less on image refinement methods and takes more advantage of light-field redundancy instead. We have shown the efficacy of our proposed refinements on three very different cost cues and thoroughly analyzed it on a publicly available robust dataset.

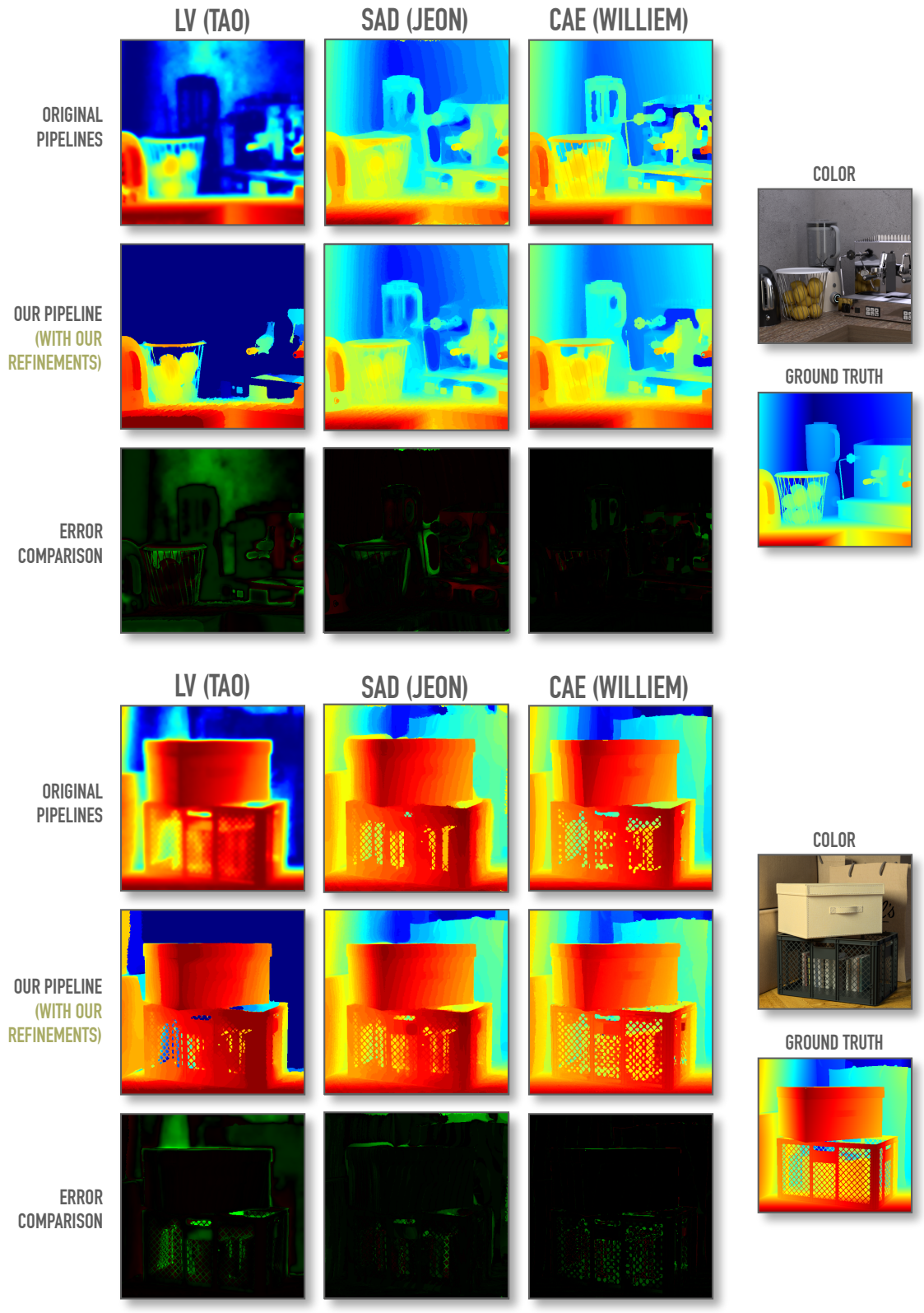


Fig. 7: Analysis of refinement performance in two scenes from synthetic dataset [28]. Top rows show depth predictions using three existing cost volume estimation methods, each processed exactly as in their original work (see Figure 2). Middle rows show predictions of the same cost volumes, but processed with our pipeline instead (which includes refinements described in Sections IV-B and IV-C). Finally, we compare the per pixel ground truth reconstruction errors between each top and middle row pair, as explained in Section V. Green means a lower error, red a higher.

Future work could improve on noise and optical effect refinement. For example, our smoothing method, when compared to image-based refinement methods used by previous work (such as weighted median transfer or iterative super resolution refinement) does not take advantage of color information, which could be a factor of improvement in the future.

REFERENCES

[1] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. So Kweon, "Accurate depth map estimation from a lenslet light field camera," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1547–1555, 2015.

[2] M. Rerabek and T. Ebrahimi, "New light field image dataset," tech. rep., 2016.

[3] T. Georgiev, Z. Yu, A. Lumsdaine, and S. Goma, "Lytro camera technology: theory, algorithms, performance analysis," in *Multimedia Content and Mobile Devices*, vol. 8667, p. 86671J, International Society for Optics and Photonics, 2013.

[4] C. Perwass and L. Wietzke, "Single lens 3d-camera with extended depth-of-field," in *Human Vision and Electronic Imaging XVII*, vol. 8291, p. 829108, International Society for Optics and Photonics, 2012.

[5] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, P. Hanrahan, *et al.*, "Light field photography with a hand-held plenoptic camera," *Computer Science Technical Report CSTR*, vol. 2, no. 11, pp. 1–11, 2005.

[6] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, "Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing," in *ACM transactions on graphics (TOG)*, vol. 26, p. 69, ACM, 2007.

[7] Z. Yu, X. Guo, H. Lin, A. Lumsdaine, and J. Yu, "Line assisted light field triangulation and stereo matching," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2792–2799, 2013.

[8] Y. Bok, H.-G. Jeon, and I. S. Kweon, "Geometric calibration of micro-lens-based light field cameras using line features," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 2, pp. 287–300, 2017.

[9] C. Shin, H.-G. Jeon, Y. Yoon, I. So Kweon, and S. Joo Kim, "Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4748–4757, 2018.

[10] W. Zhou, E. Zhou, G. Liu, L. Lin, and A. Lumsdaine, "Unsupervised monocular depth estimation from light field image," *IEEE Transactions on Image Processing*, vol. 29, pp. 1606–1617, 2019.

[11] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 673–680, 2013.

[12] M. W. Tao, P. P. Srinivasan, J. Malik, S. Rusinkiewicz, and R. Ramamoorthi, "Depth from shading, defocus, and correspondence using light-field angular coherence," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1940–1948, 2015.

[13] I. K. W. Williem, Park and K. M. Lee, "Robust light field depth estimation using occlusion-noise aware data costs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 10, pp. 2484–2497, 2018.

[14] J. Li, E. Li, Y. Chen, L. Xu, and Y. Zhang, "Bundled depth-map merging for multi-view stereo," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2769–2776, IEEE, 2010.

[15] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *International journal of computer vision*, vol. 1, no. 1, pp. 7–55, 1987.

[16] J. Mellor, S. Teller, and T. Lozano-Pérez, "Dense depth maps from epipolar images," 1996.

[17] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. S. Kweon, "Depth from a light field image with learning-based matching costs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 297–310, 2019.

[18] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *European conference on computer vision*, pp. 82–96, Springer, 2002.

[19] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[20] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2007.

[21] H. Lin, C. Chen, S. Bing Kang, and J. Yu, "Depth recovery from light field using focal stack symmetry," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3451–3459, 2015.

[22] W. N. Klarquist, W. S. Geisler, and A. C. Bovik, "Maximum-likelihood depth-from-defocus for active vision," in *Proceedings 1995 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human Robot Interaction and Cooperative Robots*, vol. 3, pp. 374–379, IEEE, 1995.

[23] M. Subbarao, T. Yuan, and J. Tyan, "Integration of defocus and focus analysis with stereo for 3d shape recovery," in *Three-Dimensional Imaging and Laser-Based Systems for Metrology and Inspection III*, vol. 3204, pp. 11–24, International Society for Optics and Photonics, 1997.

[24] H. Hirschmüller, P. R. Innocent, and J. Garibaldi, "Real-time correlation-based stereo vision with reduced border errors," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 229–246, 2002.

[25] Y. Anisimov and D. Stricker, "Fast and efficient depth map estimation from light fields," in *2017 International Conference on 3D Vision (3DV)*, pp. 337–346, IEEE, 2017.

[26] M.-J. Kim, T.-H. Oh, and I. S. Kweon, "Cost-aware depth map estimation for lytro camera," in *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 36–40, IEEE, 2014.

[27] M. Sela, E. Richardson, and R. Kimmel, "Unrestricted facial geometry reconstruction using image-to-image translation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1576–1585, 2017.

[28] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4d light fields," in *Asian Conference on Computer Vision*, pp. 19–34, Springer, 2016.