Learning Dense Wide Baseline Stereo Matching for People

Akin Caliskan[1], Armin Mustafa[1], Evren Imre[2], Adrian Hilton[1]
[1]Center for Vision, Speech and Signal Processing
University of Surrey, UK
[2] Vicon Motion Systems Ltd.
[1]{a.caliskan, a.mustafa, a.hilton}@surrey.ac.uk, [2]evren.imre@vicon.com

Abstract

Existing methods for stereo work on narrow baseline image pairs giving limited performance between wide baseline views. This paper proposes a framework to learn and estimate dense stereo for people from wide baseline image pairs. A synthetic people stereo patch dataset (S2P2) is introduced to learn wide baseline dense stereo matching for people. The proposed framework not only learns human specific features from synthetic data but also exploits pooling layer and data augmentation to adapt to real data. The network learns from the human specific stereo patches from the proposed dataset for wide-baseline stereo estimation. In addition to patch match learning, a stereo constraint is introduced in the framework to solve wide baseline stereo reconstruction of humans. Quantitative and qualitative performance evaluation against state-of-the-art methods of proposed method demonstrates improved wide baseline stereo reconstruction on challenging datasets. We show that it is possible to learn stereo matching from synthetic people dataset and improve performance on real datasets for stereo reconstruction of people from narrow and wide baseline stereo data.

1. Introduction

Recent developments in augmented reality/virtual reality and autonomous driving has led to a need for high-quality 3D content, especially for humans. However, existing scanning technologies require advanced camera setups, and controlled studio capture environments, which are complex and costly solutions. To address the need for democratization of high-quality 3D content, we propose dense stereo reconstruction for humans from wide baseline image pairs

Existing dense stereo reconstruction methods are broadly divided in two groups; narrow baseline and wide baseline stereo. For narrow baseline stereo, it is possible to estimate pixel matches by using conventional [11, 32] or learning based methods [36, 37, 14, 7]. Recently learning based methods have gained attention by outperforming conventional methods, as illustrated in the benchmark [21]. However, for wide baseline stereo, the research has focused on conventional methods [34, 12], and data driven learning based approaches are still an open research question because of the lack of training data. Results from [15, 28, 24] demonstrate that conventional wide baseline stereo methods have limitation on finding accurate matching for the human body surface. Inspired by the narrow-baseline learning based approaches and need for human specific wide baseline stereo matching, we propose a framework to estimate wide baseline dense stereo matching for people. We exploit a Siamese architecture [6] and fully connected network to learn stereo matching (Section 3.1). However existing datasets for learning stereo matching are designed for narrow baseline images with fixed relative camera locations [20, 21, 10]. In this work we introduce a synthetic human specific wide baseline stereo dataset to overcome the lim-

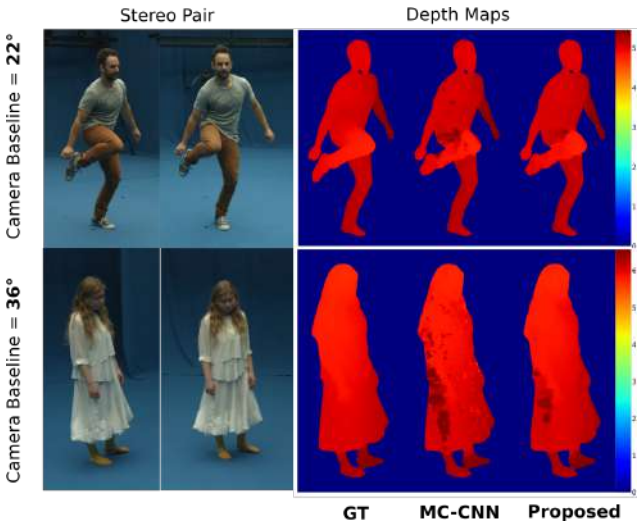

Figure 1. This figure shows the wide baseline stereo input and depth map estimation of our method with state of the art patch based stereo disparity estimation MC-CNN [37] compared to ground truth depth map.

itations of existing datasets. To refine the stereo matching performance we also propose the use of constrained stereo search using a semantic mask which is demonstrated to further refine the stereo matching performance.

Recently, the Mannequin Challenge dataset [17] has been introduced to learn monocular depth estimation for humans from a dataset of frozen people in the scene. However, this dataset does not address the problem of learning stereo matching from pairs of images. Hence, there is a need for a new dataset to perform stereo correspondence with significant change in appearance between views due to the high surcease shape of humans which includes dynamic non-rigid motion and loose clothing. Capturing a high-frequency human body with accurate ground-truth shape requires advanced scanning system which is expensive and is not easily accessible. Hence, we propose to generate stereo patch dataset for people (*S2P2*) from synthetic 3D human models with realistic textures. This dataset is used to train the stereo matching network to learn features to compare stereo image patches on the human body surface. Commonly networks trained with synthetic data do not perform well on real datasets due to the problem of domain shift [20]. Methods have tried to address this problem in the literature for different applications such as semantic segmentation [31], stereo reconstruction and optical flow estimation [20] and scene understanding [30]. These methods increase the variation in dataset by augmentation of training data with random spatial operations [5] or by creating realistic data. We have exploited these ideas by using realistic textures and applying augmentation to generated patches (Section 3.2). Another problem introduced by the domain gap is *scale diversity*. The scale of objects such as people in the image is unknown and potentially limiting the performance of a model trained on synthetic dataset. [25] proposes to use stereo pairs in various sizes to generalise training for scale diversity. Inspired by this work we extract features at different scales and combine them to get the final matching cost. We demonstrate the accuracy of the proposed stereo reconstruction on cluttered real world dataset of people in the experiments. A comprehensive performance evaluation is performed to evaluate our method with ground-truth 3D reconstruction of dynamic shape from state-of-the-art from multiple views studio performance capture. Comparison of our method with baseline methods for stereo matching shows the superior performance of features learned from the *S2P2* synthetic human dataset on wide baseline dense dynamic human stereo reconstruction. Our contributions are:

- Introduction of the first learning based framework to estimate dense wide baseline stereo for people.
- A large scale, synthetic stereo patch dataset for people with realistic textures for both narrow and wide camera baseline stereo.
- Augmentation of data and matching across multiple scales to make proposed method robust to problem of domain shift and scale.
- Refinement of learnt human stereo matching using a semantic human mask for improved stereo reconstruction.

2. Related Work

Dynamic Human Stereo Reconstruction: Existing methods for stereo reconstruction of dynamic scenes estimate correspondences between image pairs to obtain accurate surface reconstruction [28, 15] for wide baseline images. Daisy [34] and Normalized Cross Correlation (NCC) [12] uses gradient of local patch's around pixels to compute descriptor or pixel colour distribution of local patch [12] to measure patch correlation for dense wide baseline matching. In previous approaches, computation of a patch similarity measure is used as a photo-metric loss term in the objective function of optimization schema which exploits other priors, such as optical flow, edges or foreground/background segmentation [24, 28, 15]. In other words, dynamic wide baseline stereo reconstruction has not been considered as an individual solution.

Recently, learning based approaches for stereo matching have gained attention [36, 37, 14, 7] for stereo disparity estimation. However, these are trained on general scenes and are limited to narrow baseline stereo matching. To the best of our knowledge, human specific wide baseline stereo has not been addressed with learning based approaches before. Previous work [16] trains a multi-view patch similarity network for performance capture using the DTU general object dataset [3]. This paper addresses this gap in the literature, by proposing dense stereo reconstruction from wide baseline image pairs and learning to perform stereo matching using a new synthetic people stereo patch dataset (*S2P2*).

Learning Depth from Synthetic Data Recently, usage of synthetic data to train neural networks for depth estimation has gained attention. One of the first synthetic data-set proposed is [20]. This data is used to train a network for narrow baseline stereo disparity and optical flow estimation. Another work from [10] generates the virtual version of Kitti data-set [21]. Virtual Kitti includes additional annotations like segmentation, depth estimation and 3D object tracking. They demonstrate that training on synthetic data and using learned model on real data is possible. Varol *et al.* [35] proposed a synthetic human dataset for monocular model based human segmentation and depth estimation. However, synthetic data trained models suffer from limitations on real world images in high-frequency depth estimation of the human body [29]. [13] introduced another synthetic human dataset to train multi-view surface estimation network. We propose a large scale stereo patch dataset (S2P2) for people to train a network for wide baseline dense stereo matching across difference scales. To the best of our knowledge, the

proposed *(*S2P2) dataset is the first to learn stereo matching for people. This dataset can be used for both narrow and wide baseline stereo estimation.

3. Method

The main motivation of this work is to estimate 3D reconstruction of humans in dynamic scenes from wide baseline stereo camera pairs. Note that, the difference between narrow ($\theta <= 5^\circ$) and wide baseline ($15^\circ < \theta < 45^\circ$) cameras is illustrated in Figure 2 - (a). We propose a supervised learning based framework which first learns stereo matching from a new synthetic human specific dataset *S2P2* for wide-baseline cameras followed by stereo reconstruction refinement using semantic human constraint, an overview is illustrated in Figure 2 - (b). Variation of human body surface for example folded clothing, hair, face details, makes it challenging to extract reliable stereo reconstruction from wide baseline image pairs. Given a wide baseline stereo pair of images of a person, we aim to obtain per-pixel dense correspondence for stereo reconstruction. The stereo pair of images are fed into a CNN module, which is trained on a human specific dataset to obtain the matching cost for each pixel. This generates a cost volume which is refined using a semantic stereo constraint to obtain the final depth map. In the following sections, the patch match learning architecture (Section 3.1), data generation pipeline (Section 3.2), the method to solve domain shift from synthetic to real data (Section 3.2) and semantic stereo constraint (Section 3.3) are explained in detail.

3.1. Learning Wide Baseline Stereo Matching

The overall CNN module for learning stereo matching is illustrated in Figure 2 - (c). We use a Siamese network architecture [6] as the backbone, which has received a lot of attention lately for various applications including patch based binary classification [37], and patch based tracking [4]. Siamese network is suitable for the proposed application because it allows training of stereo matching between a pair of left and right image patches. Methods used Siamese network as feature extraction module in patch based narrow baseline stereo matching [19, 37]. The network consists of four consecutive 2D convolution layers and RELU (Rectified Linear Units) after each convolution layer. As illustrated in Figure 2 - (c), the computed feature vectors are fed into a fully-connected network (FCN) to estimate the similarity score between patches, i.e. classification module. The details of CNN module is provided in the supplementary file. Since we are solving a binary classification problem, we use binary cross entropy loss [22] to train our network. During the training stage, we use a balanced number of positive and negative patches extracted from the *S2P2* dataset (Sec. 3.2).

In the implementation stage, multi resolution patches are extracted for each pixel followed by resizing the patches to a fixed patch size that the network is trained with. Patches are processed through the network, and matching cost is computed for each patch. Individually generated cost volumes are fed into the pooling stage as illustrated in Figure 2, where the resultant matching cost is computed from the similarity scores for each pixel pairs. In the pooling stage, the matching cost from different patch sizes are gathered and the average value is assigned as a final cost. We evaluate the effect of pooling by comparing the results with or without pooling in the Experiment section - Table 5.

We compute the cost volume for both left and right camera views respectively and a winner takes all method is applied to each of the views to compute the final disparity values. In contrast to the conventional stereo pipelines [32, 37] which require heavy regularization steps for post-processing like Semi-Global Matching [11] and Bilateral filtering [32], we perform a simple post-processing to remove the occlusions on the estimated disparity maps to improve stereo from wide baseline image pairs.

3.2. Synthetic People Stereo Patch Dataset

Existing datasets in the literature are limited to narrow-baseline general scenes. We address this gap in the literature by proposing a data generation framework for supervised wide baseline stereo matching learning for people, illustrated in Figure 2. We generate the dataset by using the blender 3D modelling[1]. Parametric 3D SMPL [18] human models are generated based on 3D pose estimation from real humans with random shape parameters, CMU MoCap Dataset [1]. Then realistic textures are rendered on the generated models. Up to this point, model generation is inspired by the Surreal dataset [35].

To add varied backgrounds to each image a 3D plane is placed behind the person model and background scene images are randomly selected from Places Database [38], which consists of high variation of indoor and outdoor places with different configurations. Camera locations and orientations are replicated from real studio capture setups, and the baseline between cameras is varied from narrow (5°) to wide (up to 45° degrees). The generated scene is then rendered into camera views with random lighting settings. For training purposes, we generate patches from non-occluded regions of the human body surface.

Proposed network structure requires positive and negative patches. Positive patches are generated from projection of 3D points into stereo views and negative patches are ψ pixels away from positive patches along the Epipolar line. During training data generation, patch size is fixed to 9x9, and ψ value is randomly selected from interval $[4, 11]$. Reference patch with positive and negative pairs are augmented [5] in spatial and spectral domains which includes random

[1] https://www.blender.org

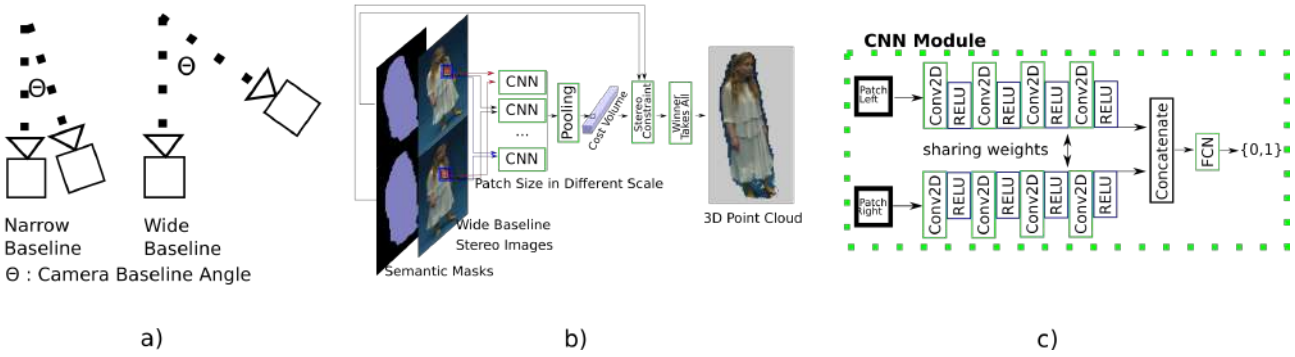

Figure 2. (a) Definition of camera baseline angle, θ and difference between narrow and wide camera baselines.(b) The proposed stereo reconstruction method.(c) CNN module used for patch match learning part of the proposed method.

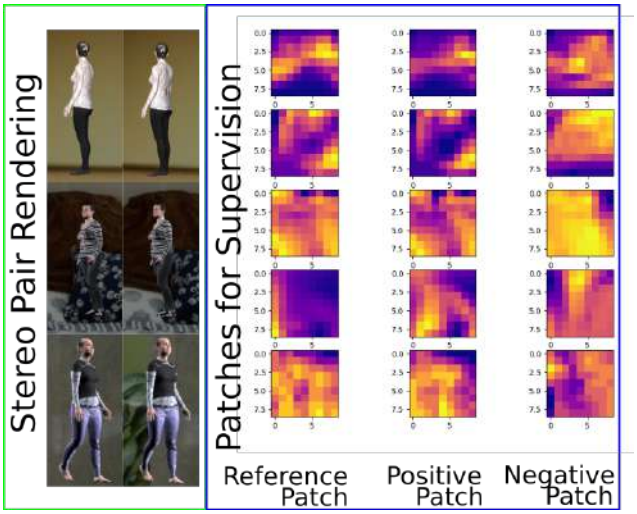

Figure 3. **Data Generation Pipeline:** SMPL human model [18] is generated with random shape, texture and given 3D pose parameters. Cameras are placed according to real studio calibration in different baseline configuration. [*Left*] Rendering of human models onto camera planes with different background and 3D pose. [*Right*] Positive and negative patches are generated from these images. For details, please refer to text.

cropping, flipping, transformation, and contrast variation. This dataset along with the data generation framework is available for public use [2] and further details of the dataset are given in supplementary material.

Data Augmentation and Scale Invariance: Learning from the synthetic dataset and testing on real images has recently gained attention in the literature for different applications [20, 10, 35, 27]. The common problem is domain adaptation which directly affects the learning from synthetic to real imagery. In our work, we generate *S2P2* dataset from a wide variety of camera positions and realistically textured human models. We add patch augmentation, explained previously, to increase the robustness of stereo correspondence for real data. However real data can be observed with different input scale than synthetic data, which results in stereo correspondence defects, called *scale diversity* [25]. Since the scale of real data is unknown, we look for the consistency of accurate matches for different patch sizes before computation of the final cost volume in the *pooling stage*, illustrated in Figure 2 - (b). To address this, we take the average of matching cost values that are computed with the trained network for every pixel. This multi-scale patch size approach is analyzed for real data and Table 5 shows the performance improvement in the reconstruction accuracy.

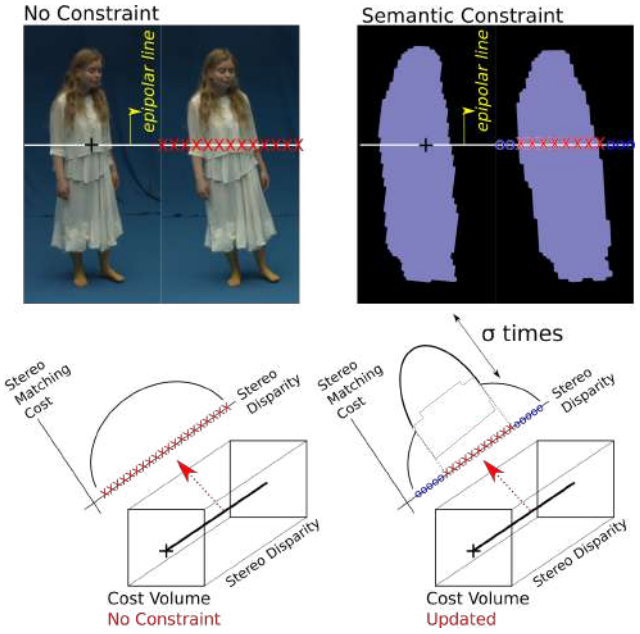

Figure 4. Semantic mask based stereo constraint for wide baseline stereo on Acting dataset [2].

3.3. Semantic Stereo Constraints

To further refine the learnt stereo matching we introduce a semantic stereo constraint for stereo matching on people leveraging recent advances in semantic segmentation. Stereo matching requires reliable per-pixel correspon-

[2] https://akcalakcal.github.io/Learning-Dense-Wide-Baseline-Stereo-Matching-for-People/

dences in between image pairs. With a given calibration, patch match methods rectify the images to find correspondences along the epipolar line by comparing the pixel similarities. However due to drastic view variation in wide baseline stereo pairs, patch match methods fail to find reliable correspondences. Previous studies on wide baseline human performance capture methods [24, 33] either use initial sparse reconstruction or visual hulls generated from multi cameras to limit the stereo search space. Other methods for wide baseline semantic reconstruction exploit semantic segmentation constraints to improve the multi-view stereo [23]. In this study, we propose to exploit semantic masking in the stereo matching framework to limit the search region along the Epipolar line to decrease the number of wrong matches from only two camera views. However errors in semantic segmentation do not adversely affect the accuracy of the reconstruction, unlike previous method.

We use DeepLabv3+ [8] to obtain the semantic masks. The correspondence search algorithm for two stereo rectified images and corresponding semantic masks is illustrated in Figure 4. Without constraint, a pixel in the left image is compared with all the pixels in the corresponding right image. However, with the semantic constraint we search for the corresponding pixel within the semantic region along the Epipolar line reducing the ambiguity and run-time complexity. The cost volume in Figure 4 is processed with the semantic constraint such that for pixels in the masked region, the cost value is weighted by a coefficient, $\sigma = 10$. This suppress other pixels for matching.

4. Experiments

We answer the following questions in experiments:

- Does learning wide baseline stereo matching from people dataset result in better matching for image pairs of people that existing approaches which learn from non-human stereo dataset and conventional methods?
- Does the proposed solution to domain shift with patch augmentation and scale diversity, improve the reconstruction results for real datasets with humans?
- Does the proposed semantic human stereo constraint improve the stereo reconstruction results?

Implementation and Training Details The network architecture is implemented in PyTorch [26] framework on a single NVIDIA GeForce 1080 Ti GPU with 12 GB memory. As described in Section 3.1, we train our model from scratch. The learning rate is initialized at 3×10^{-3} with a 10 times decrease at every 10 epochs. Training is performed for 15 epochs, and momentum and weight decay are set to 0.9 and 0.0001, respectively. The entire network is learned with stochastic gradient descent optimization with binary cross entropy loss function. The network weights are randomly initialized with balanced number of positive and negative patches with a total of 14 million patches. Variation of training loss versus epochs is illustrated in Figure 5.

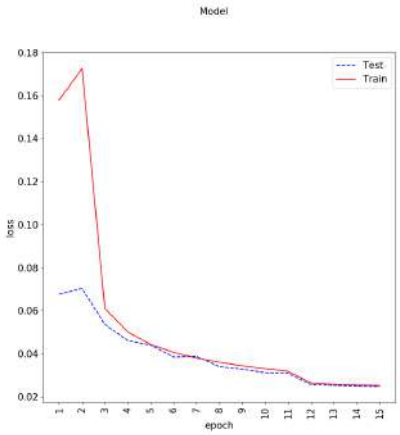

Figure 5. Variation of loss during training.

4.1. Results and Comparisons

The proposed method is evaluated on a variety of real datasets with people in different environments, cluttered/controlled scene background, occlusions, camera settings and baselines: Acting, TV-Presenter, Dancing and Juggler [2]. The details of datasets are provided in Table 1. These datasets consist of different dynamic human models and each scene is captured with number of cameras given in the Table. In these datasets, we use pseudo ground truth of 3D human reconstructions that are generated by using advanced multi-view camera capture system. For each camera view, ground-truth depth maps are rendered and then estimated 3D stereo reconstructions are evaluated against these rendered depth maps. Synthetic datasets for testing are different from training datasets and are generated using the framework explained in Section 3.2.

The proposed method is evaluated against baseline patch matching methods, namely NCC [12], Daisy [34] and MC-CNN [37] since we propose a patch similarity based wide baseline stereo reconstruction method. MC-CNN is a state-of-the-art baseline method for stereo matching. MC-CNN is built on a Siamese network architecture and this network is trained on the Kitti [21] dataset of narrow baseline stereo street images taken from top of the car with sparse ground-truth obtained by lidar scanner.

We adopt the following error metrics [9] to quantitatively evaluate the performance of our stereo reconstruction

Dataset	Camera Baseline	# of Cameras	Length of Sequence (# of Frames)
Acting	{24°,36°,48°}	15	3420
TV Presenter	{22°,44°,66°}	16	3600
Dancing	{22°,44°,66°}	16	420
Juggler	{22°,44°,66°}	8	800

Table 1. Real World of People Datasets.

	Camera Baseline $\approx 20^{\circ}$				Camera Baseline $\approx 40^{\circ}$			
Method	Abs Rel	Squ Rel	RMSE	RMSE_{log}	Abs Rel	Squ Rel	RMSE	RMSE_{log}
Dataset:Acting								
NCC [12]	3.40	3.18	44.3	7.89	5.21	3.52	46.6	8.10
Daisy [34]	1.77	0.92	24.0	3.53	2.05	0.95	24.1	3.70
MC-CNN [37]	1.27	0.11	8.56	1.36	1.42	0.43	16.1	2.57
Ours	**0.70**	**0.04**	**5.30**	**0.86**	**1.03**	**0.26**	**12.6**	**1.99**
Dataset:Dancing								
NCC [12]	6.78	4.89	49.9	10.6	6.08	2.51	35.3	7.29
Daisy [34]	1.83	0.75	19.4	3.74	2.55	0.88	20.6	3.89
MC-CNN [37]	1.12	0.38	13.8	2.52	1.76	0.39	17.3	3.41
Ours	**0.84**	**0.16**	**8.69**	**1.68**	**1.71**	**0.33**	**15.3**	**3.01**

Table 2. Depth estimation error results for 2 datasets against four compared methods are listed in the table. For details of experiment and error metrics, please refer to text.

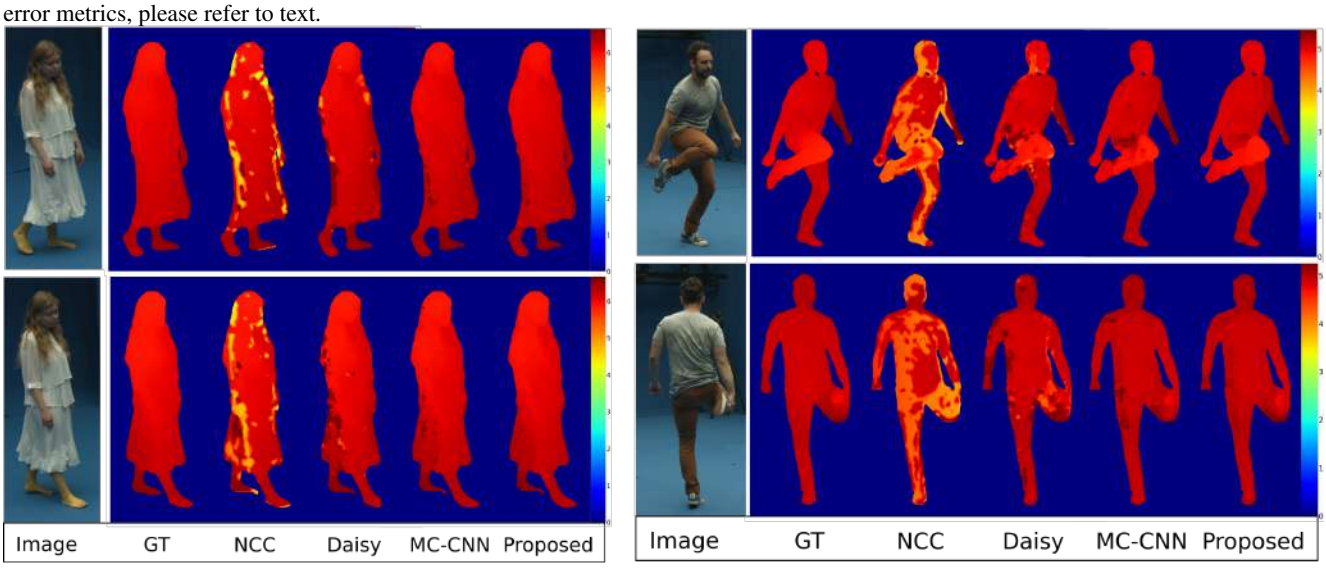

Figure 6. Comparison of estimated depth maps with ground-truth. Result depth maps of four methods, namely NCC [12], Daisy [34] and MC-CNN [37], including proposed one are illustrated. Camera baseline between stereo pairs are 24°.

method. Established error metrics consider global statistics between a predicted depth map d and its ground-truth depth image d^* with N depth pixels. Specifically, we consider: (i) *absolute relative error*: $\frac{1}{N}\sum_i \frac{|d_i - {d_i}^*|}{{d_i}^*}$; (ii) *squared relative error*: $\frac{1}{N}\sum_i \frac{||d_i - {d_i}^*||^2}{{d_i}^*}$; (iii) *root mean square error*: $\sqrt{\frac{1}{N}\sum_i (d_i - {d_i}^*)^2}$; (iv) *logarithmic root mean square error*: $\sqrt{\frac{1}{N}\sum_i (\log d_i - \log {d_i}^*)^2}$.

	Lower is better			
Method	Abs Rel	Squ Rel	RMSE	RMSE_{log}
Dataset:Acting				
MC-CNN [37]	1.27	0.11	8.56	1.36
MC-CNN [37] w/ constraint	0.76	0.08	7.30	1.16
Ours	**0.70**	**0.04**	**5.30**	**0.86**
Dataset:Dancing				
MC-CNN [37]	1.12	0.38	13.8	2.52
MC-CNN [37] w/ constraint	0.98	0.31	12.3	2.41
Ours	**0.84**	**0.16**	**8.69**	**1.68**

Table 3. Depth map evaluation with and without stereo constraint.

Table 2 shows depth error metrics for two different datasets with two different wide baselines. For this experiment, baseline between stereo pairs is 24° and 36° for *Acting*, 22° and 44° for *Dancing* datasets. Corresponding depth estimation results are illustrated with ground-truth (GT) depth maps in Figure 6. As shown in Table 2, the proposed method outperforms the baseline methods in terms of depth map estimation errors for wide baseline datasets. The proposed method gives approximately 25% RMSE er-

ror reduction for two camera baseline values compared to MC-CNN, which is the state of the art patch based stereo reconstruction method. It should also be considered that MC-CNN applies a series of expensive post processing steps, like occlusion removal, Semi-Global-Matching (SGM) [11] and Bilateral filtering, where as proposed method only applies occlusion removal and not any of smoothing operations to recover wrong disparity estimations. Considering these post processing steps, for the same input stereo pairs with resolution of $3840x2160$ pixels, the run time for MC-CNN is 210 seconds whereas the proposed method only takes 135 seconds. Hence, the proposed method not only outperforms MC-CNN in depth error metrics, but also it is faster than MC-CNN by approximately 35%.

Figure 7 shows the point clouds and depth maps, demonstrating a significant difference between the proposed method and MC-CNN. Depth values in the GT depth maps are defined in meters. Note that during the depth map error computation, only the foreground pixels are evaluated, and background pixels are discarded.

The proposed method also outperforms NCC [12] and Daisy [34] in all depth estimation metrics. NCC [12] and Daisy [34] generate local descriptors that are prone to fail in ambiguities, like repetitive textures, lack of textures, or lighting changes and large changes in shape. These failures can be resolved during post processing stage in wide baseline human stereo reconstruction methods [28, 15, 24].

The reconstruction results are shown in Figure 7 with corresponding depth maps for MC-CNN and the proposed method. In addition to depth error metrics, 3D point clouds show the details in reconstruction. In Figure 7, dynamic 3D stereo reconstruction of human body is also illustrated for different time frames. The generated point clouds are rendered to virtual cameras in order to see the stereo reconstruction errors that might be difficult to see from depth maps. The proposed method which learns from human specific features is able to capture details of clothing and hair which are challenging to reconstruct in wide baseline stereo setups. This answers the first question, that learning from a human-specific dataset improves wide baseline stereo performance.

Another contribution of our paper is to use semantic segmentation based stereo limitation to improve stereo matching performance or the reconstruction quality (Section 3.3). This constraint can be applied to any stereo matching method, so we evaluate the stereo matching performance of state-of-the-art methods with this constraint. During evaluation, only MC-CNN and the proposed stereo matching method are considered, because remaining methods' stereo reconstruction performance is not affected significantly with the constraint. In Table 3, semantic constrained is applied to MC-CNN for different datasets. Although semantic constraint increases the performance of

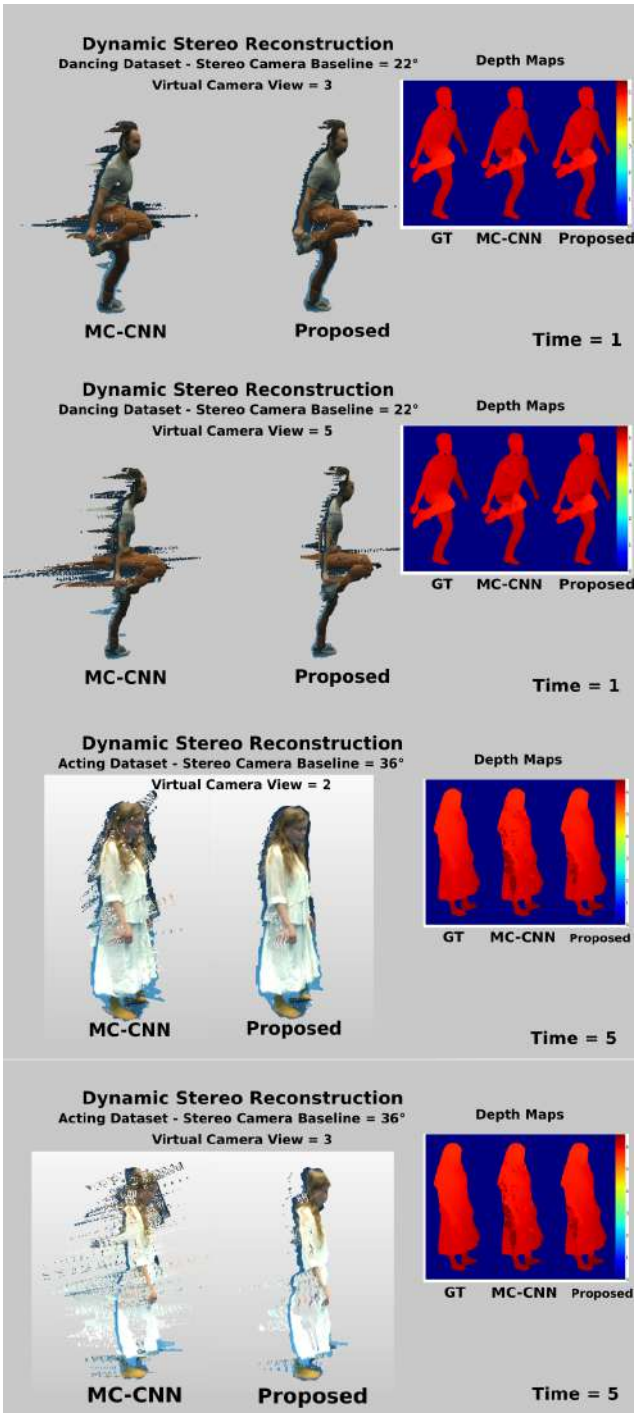

Figure 7. Point cloud stereo reconstruction results with depth map estimations from various time frames are illustrated for virtual camera views.

MC-CNN by approximately 12% in RMSE, the proposed method still outperforms MC-CNN with stereo constraint in all error metrics, by average of $\%30$ in RMSE.

To evaluate the importance of the new *S2P2* dataset, we evaluate performance of patch matching part of the pro-

Method	Lower is better			
	Abs Rel	Squ Rel	RMSE	$RMSE_{log}$
Dataset:Acting				
Our Method w/ Kitti Dataset	0.85	0.12	8.7	1.41
Our Method w/ ($S2P2$) no augmentation	0.67	0.09	7.8	1.25
Our Method w/ ($S2P2$) Dataset	**0.63**	**0.07**	**6.93**	**1.07**
Dataset:Dancing				
Our Method w/ Kitti Dataset	1.60	0.38	13.7	2.67
Our Method w/ ($S2P2$) no augmentation	1.22	0.39	14.0	2.56
Our Method w/ ($S2P2$) Dataset	**0.81**	**0.11**	**7.68**	**1.52**

Table 4. Dataset and domain shift evaluation

posed framework with two different models one of which is trained with *S2P2* dataset, and other one is trained with Kitti dataset, shown in Table 4. Our method using the *S2P2* trained network outperforms the network trained on Kitti, by approximately 30% in logarithmic RMSE. This basically shows that learning stereo matching from wide baseline and human specific data in our framework addresses more accurate wide baseline stereo reconstruction for people, which is the motivation of this paper. Table 4 also demonstrates that data augmentation on stereo people dataset improves accuracy of depth maps and addresses the problem of domain shift from training on synthetic data and testing on real data.

As a part of our solution to scale variance in our method, we propose the pooling schema during inference stage of stereo reconstruction. In the pooling, we use patch size values of [9,19,35] in order to increase the patch scale variation. In order to show the effectiveness of the pooling stage, we evaluate proposed method with and without pooling and compare the results with MC-CNN [37]. Since patch size is chosen as 9x9 in [37], we use this patch size during no-pooling evaluation. Depth estimation errors in Table 5 demonstrate that pooling stage in the pipeline increases accuracy of stereo reconstruction by solving scale diversity problem caused by domain shift.

To illustrate the performance of proposed method with human in dynamic scene with cluttered background, point cloud results are shown on *Juggler* dataset for consecutive frames in Figure 8. Note that stereo input images of juggler dataset are cropped for the visualization. The estimated point clouds from both front and side views show significant reconstruction performance from the proposed wide baseline stereo matching method with semantic human constraint. More stereo reconstruction results from both real and synthetic datasets are provided in supplementary files due to space constraint.

5. Limitations

The proposed method is developed for wide baseline stereo reconstruction for people, and this is not applicable

Method	Lower is better			
	Abs Rel	Squ Rel	RMSE	$RMSE_{log}$
Dataset:TV Presenter, Patch Size = (9x9)				
MC-CNN [37]	0.67	0.07	6.18	1.13
Our method + No Pooling	0.61	0.06	5.94	1.08
Our Method + Pooling	**0.60**	**0.05**	**5.33**	**0.98**

Table 5. Scale Diversity Evaluation

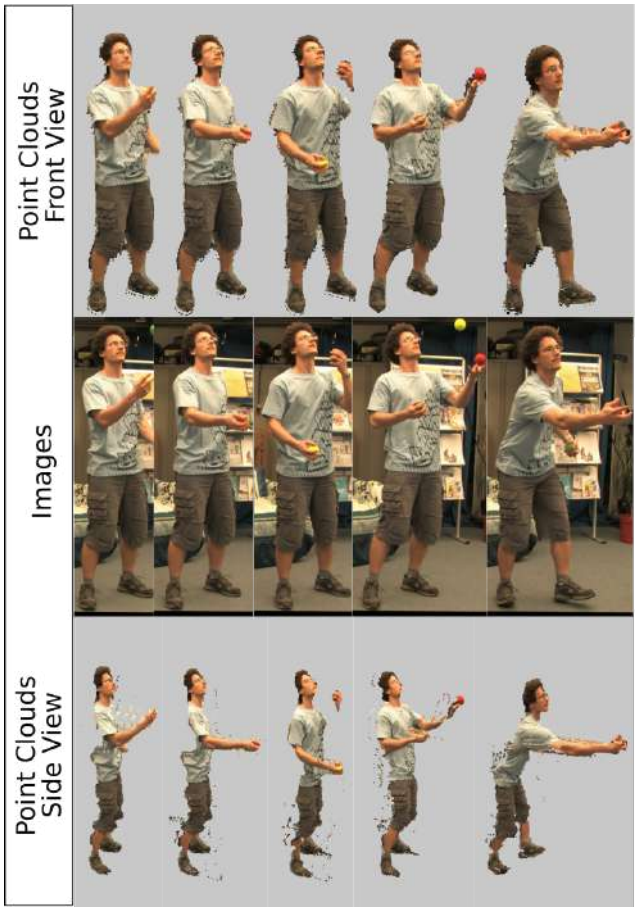

Figure 8. This figure illustrates the wide baseline stereo reconstruction of human in dynamic scene with cluttered background and 22° camera baseline angle.

to solve wide baseline stereo for generic scenes. However, a supervised learning based method for generic scenes is possible with provided training data and whole scene segmentation.

6. Conclusion

In this paper we proposed a method to solve the challenging task of wide baseline dense stereo reconstruction of humans. A framework to learn human specific features for stereo reconstruction from synthetic people stereo patch dataset is introduced. Multiple patch sizes are used to extract features and fused using pooling to address the problem of adapting the network from synthetic to real data. Comparative performance evaluation demonstrates that the learnt stereo matching outperforms state-of-the-art methods in human reconstruction and is robust to wide baseline and scale changes. To further refine the stereo reconstruction a person specific semantic stereo matching constraint is introduced. Extensive performance evaluation on real datasets shows that the proposed method outperforms state-of-the-art methods.

References

[1] Carnegie-mellon graphics lab motion capture database. http://http://mocap.cs.cmu.edu. Accessed: 2019-06-05.

[2] Multiview video repository,. http://cvssp.org/data/cvssp3d/. In Centre for Vision Speech and Signal Processing, University of Surrey, UK.

[3] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016.

[4] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016.

[5] M. D. Bloice, C. Stocker, and A. Holzinger. Augmentor: an image augmentation library for machine learning. *arXiv preprint arXiv:1708.04680*, 2017.

[6] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.

[7] R. Chabra, J. Straub, C. Sweeny, R. Newcombe, and H. Fuchs. Stereodrnet: Dilated residual stereo net. *arXiv preprint arXiv:1904.02251*, 2019.

[8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

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

[10] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016.

[11] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007.

[12] X. Hu and P. Mordohai. A quantitative evaluation of confidence measures for stereo vision. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2121–2133, 2012.

[13] Z. Huang, T. Li, W. Chen, Y. Zhao, J. Xing, C. LeGendre, L. Luo, C. Ma, and H. Li. Deep volumetric video from very sparse multi-view performance capture. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 336–354, 2018.

[14] A. Kendall, H. Martirosyan, S. Dasgupta, and P. Henry. End-to-end learning of geometry and context for deep stereo regression. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 66–75. IEEE, 2017.

[15] V. Leroy, J.-S. Franco, and E. Boyer. Multi-view dynamic shape refinement using local temporal integration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3094–3103, 2017.

[16] V. Leroy, J.-S. Franco, and E. Boyer. Shape reconstruction using volume sweeping and learned photoconsistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 781–796, 2018.

[17] Z. Li, T. Dekel, F. Cole, R. Tucker, N. Snavely, C. Liu, and W. T. Freeman. Learning the depths of moving people by watching frozen people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4521–4530, 2019.

[18] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015.

[19] W. Luo, A. G. Schwing, and R. Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5695–5703, 2016.

[20] N. Mayer, E. Ilg, P. Fischer, C. Hazirbas, D. Cremers, A. Dosovitskiy, and T. Brox. What makes good synthetic training data for learning disparity and optical flow estimation? *International Journal of Computer Vision*, 126(9):942–960, 2018.

[21] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015.

[22] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[23] A. Mustafa and A. Hilton. Semantically coherent co-segmentation and reconstruction of dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 422–431, 2017.

[24] A. Mustafa, H. Kim, J.-Y. Guillemaut, and A. Hilton. General dynamic scene reconstruction from multiple view video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 900–908, 2015.

[25] J. Pang, W. Sun, C. Yang, J. Ren, R. Xiao, J. Zeng, and L. Lin. Zoom and learn: Generalizing deep stereo matching to novel domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2070–2079, 2018.

[26] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[27] M. Rad, M. Oberweger, and V. Lepetit. Feature mapping for learning fast and accurate 3d pose inference from synthetic images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4663–4672, 2018.

[28] C. Richardt, H. Kim, L. Valgaerts, and C. Theobalt. Dense wide-baseline scene flow from two handheld video cameras. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 276–285. IEEE, 2016.

[29] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019.

[30] C. Sakaridis, D. Dai, S. Hecker, and L. Van Gool. Model adaptation with synthetic and real data for semantic dense

foggy scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 687–704, 2018.

[31] S. Sankaranarayanan, Y. Balaji, A. Jain, S. Nam Lim, and R. Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3752–3761, 2018.

[32] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002.

[33] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3):21–31, May 2007.

[34] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):815–830, 2009.

[35] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, 2017.

[36] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4353–4361, 2015.

[37] J. Zbontar, Y. LeCun, et al. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1-32):2, 2016.

[38] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.