International Journal of Computer Vision (2020) 128:873–890
https://doi.org/10.1007/s11263-019-01279-w

Learning Multi-human Optical Flow

Anurag Ranjan[1] · David T. Hoffmann[1] · Dimitrios Tzionas[1] · Siyu Tang[1] · Javier Romero[2] · Michael J. Black[1]

Received: 20 January 2019 / Accepted: 10 December 2019 / Published online: 2 January 2020

Abstract
The optical flow of humans is well known to be useful for the analysis of human action. Recent optical flow methods focus on training deep networks to approach the problem. However, the training data used by them does not cover the domain of human motion. Therefore, we develop a dataset of multi-human optical flow and train optical flow networks on this dataset. We use a 3D model of the human body and motion capture data to synthesize realistic flow fields in both single- and multi-person images. We then train optical flow networks to estimate human flow fields from pairs of images. We demonstrate that our trained networks are more accurate than a wide range of top methods on held-out test data and that they can generalize well to real image sequences. The code, trained models and the dataset are available for research.

Keywords Optical · Flow · Deep · Learning · Human · Bodies · Synthetic · Dataset · Humanflow

1 Introduction

A significant fraction of videos on the Internet contain people moving (Geman and Geman 2016) and the literature suggests that optical flow plays an important role in understanding human action (Jhuang et al. 2013; Soomro et al. 2012). Several action recognition datasets (Soomro et al. 2012; Kuehne et al. 2011) contain human motion as a major component. The 2D motion of humans in video, or *human optical flow*, is an important feature that provides a building block for systems that can understand and interact with humans. Human optical flow is useful for various applications including analyzing pedestrians in road sequences, motion-controlled gaming, activity recognition, human pose estimation, etc.

Despite this, optical flow has previously been treated as a generic, low-level, vision problem. Given the importance of people, and the value of optical flow in understanding them, we develop a dataset and trained models that are specifically tailored to humans and their motion. Such motions are nontrivial since humans are complex, articulated objects that vary in shape, size and appearance. They move quickly, adopt a wide range of poses, and self-occlude or occlude in multi-person scenarios.

Our goal is to obtain more accurate 2D motion estimates for human bodies by training a flow algorithm specifically for human movement. To do so, we create a large and realistic dataset of humans moving in virtual worlds with ground truth optical flow (Fig. 1a), called the *Human Optical Flow* dataset. This is comprised of two parts; the *Single-Human Optical Flow* dataset (SHOF), where the image sequences contain only one person in motion and the *Multi-Human Optical Flow* dataset (MHOF) where images contain multiple peo-

Communicated by Ling Shao, Hubert P. H. Shum, Timothy Hospedales.

Anurag Ranjan and David T. Hoffmann have contributed equally to this work.

This work was done when JR was at MPI-IS. MJB has received research gift funds from Intel, Nvidia, Adobe, Facebook, and Amazon. While MJB is a part-time employee of Amazon, his research was performed solely at, and funded solely by, MPI. MJB has financial interests in Amazon and Meshcapade GmbH.

✉ Anurag Ranjan
anurag.ranjan@tue.mpg.de

David T. Hoffmann
david.hoffmann@tue.mpg.de

Dimitrios Tzionas
dimitris.tzionas@tue.mpg.de

Siyu Tang
siyu.tang@tue.mpg.de

Javier Romero
javier@amazon.com

Michael J. Black
black@tue.mpg.de

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany

[2] Amazon Body Labs, Barcelona, Spain

ple involving significant occlusion between them. We analyse the performance of SPyNet (Ranjan and Black 2017) and PWC-Net (Sun et al. 2018) by training (fine-tuning) them on both the SHOF and MHOF dataset. We observe that the optical flow performance of the networks improves on sequences containing human scenes, both qualitatively and quantitatively. Furthermore we show that the trained networks generalize to real video sequences (Fig. 1c). Several datasets and benchmarks (Baker et al. 2011; Geiger et al. 2012; Butler et al. 2012) have been established to drive the progress in optical flow. We argue that these datasets are insufficient for the task of human motion estimation and, despite its importance, no attention has been paid to datasets and models for human optical flow. One of the main reasons is that dense human motion is extremely difficult to capture accurately in real scenes. Without ground truth, there has been little work focused specifically on estimating human optical flow. To advance research on this problem, the community needs a dataset tailored to human optical flow.

A key observation is that recent work has shown that optical flow methods trained on synthetic data (Ranjan and Black 2017; Dosovitskiy et al. 2015; Ilg et al. 2016) generalize relatively well to real data. Additionally, these methods obtain state-of-the-art results with increased realism of the training data (Mayer et al. 2016; Gaidon et al. 2016). This motivates our effort to create a dataset designed for human motion.

To that end, we use the SMPL (Bogo et al. 2016) and SMPL+H (Romero et al. 2017) models, that capture the human body alone and the body together with articulated hands respectively, to generate different human shapes including hand and finger motion. We then place humans on random indoor backgrounds and simulate human activities like running, walking, dancing etc. using motion capture data (Loper et al. 2014; Mahmood et al. 2019). Thus, we create a large virtual dataset that captures the statistics of natural human motion in multi-person scenarios. We then train optical flow networks on this dataset and evaluate their performance for estimating human motion. While the dataset can be used to train any flow method, we focus specifically on networks based on spatial pyramids, namely SpyNet (Ranjan and Black 2017) and PWC-Net (Sun et al. 2018), because they are compact and computationally efficient.

A preliminary version of this work appeared in Ranjan et al. (2018) that presented a dataset and model for human optical flow for the *single-person* case with a *body-only* model. The present work extends (Ranjan et al. 2018) for the *multi-person* case, as images with multiple occluding people have different statistics. It further employs a holistic model of the *body together with hands* for more realistic motion variation. This work also extends training SPyNet (Ranjan and Black 2017) and PWC-Net (Sun et al. 2018) using the new dataset in contrast to training only SPyNet in the earlier work Ranjan et al. (2018). Our experiments show both qualitative and quantitative improvements.

In summary, our major contributions in this extended work are: (1) We provide the *Single-Human Optical Flow* dataset (SHOF) of human bodies in motion with realistic textures and backgrounds, having $146,020$ frame pairs for single-person scenarios. (2) We provide the *Multi-Human Optical Flow* dataset (MHOF), with $111,312$ frame pairs of multiple human bodies in motion, with improved textures and realistic visual occlusions, but without (self-)collisions or intersections of body meshes. These two datasets together comprise the *Human Optical Flow* dataset. (3) We fine-tune SPyNet (Ranjan et al. 2018) on SHOF and show that its performance improves by about 43% (over the initial SPyNet), while it also outperforms existing state of the art by about 30%. Furthermore, we fine-tune SPyNet and PWC-Net on MHOF and observe improvements of $10-20\%$ (over the initial SPyNet and PWC-Net). Compared to existing state of the art, improvements are particularly high for human regions. After masking out the background, we observe improvements of up to 13% for human pixels. (4) We provide the dataset files, dataset rendering code, training code and trained models[1] for research purposes.

2 Related Work

2.1 Human Motion

Human motion can be understood from 2D motion. Early work focused on the movement of 2D joint locations (Johansson 1973) or simple motion history images (Davis 2001). Optical flow is also a useful cue. Black et al. (1997) use principal component analysis (PCA) to parametrize human motion but use noisy flow computed from image sequences for training data. More similar to us, Fablet and Black (2002) use a 3D articulated body model and motion capture data to project 3D body motion into 2D optical flow. They then learn a view-based PCA model of the flow fields. We use a more realistic body model to generate a large dataset and use this to train a CNN to directly estimate dense human flow from images.

Only a few works in pose estimation have exploited human motion and, in particular, several methods (Fragkiadaki et al. 2013; Zuffi et al. 2013) use optical flow constraints to improve 2D human pose estimation in videos. Similar work (Pfister et al. 2015; Charles et al. 2016) propagates pose results temporally using optical flow to encourage time consistency of the estimated bodies. Apart from its application in warping between frames, the structural information existing in optical flow alone has been used for pose estima-

[1] https://humanflow.is.tue.mpg.de.

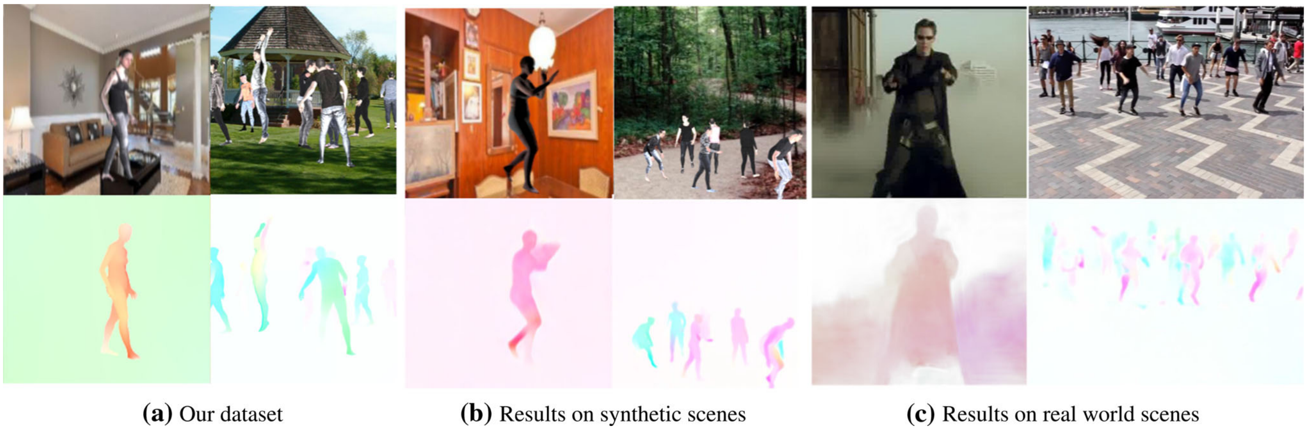
(a) Our dataset
(b) Results on synthetic scenes
(c) Results on real world scenes

Fig. 1 **a** We simulate human motion in virtual worlds creating an extensive dataset with images (top row) and flow fields (bottom row); color coding from Baker et al. (2011). **b** We train SPyNet (Ranjan and Black 2017) and PWC-Net (Sun et al. 2018) for human motion estimation and show that they perform better when trained on our dataset and **c** can generalize to human motions in real world scenes. Columns show single-person and multi-person cases alternately.

tion (Romero et al. 2015) or in conjunction with an image stream (Feichtenhofer et al. 2016; Dong et al. 2018).

2.2 Learning Optical Flow

There is a long history of optical flow estimation, which we do not review here. Instead, we focus on the relatively recent literature on learning flow. Early work looked at learning flow using Markov Random Fields (Freeman et al. 2000), PCA (Wulff and Black 2015), or shallow convolutional models (Sun et al. 2008). Other methods also combine learning with traditional approaches, formulating flow as a discrete (Güney and Geiger 2016) or continuous (Revaud et al. 2015) optimization problem.

The most recent methods employ large datasets to estimate optical flow using deep neural networks. Voxel2Voxel (Tran et al. 2016) is based on volumetric convolutions to predict optical flow using 16 frames simultaneously but does not perform well on benchmarks. Other methods (Ranjan and Black 2017; Dosovitskiy et al. 2015; Ilg et al. 2016) compute two frame optical flow using an end-to-end deep learning approach. FlowNet (Dosovitskiy et al. 2015) uses the Flying Chairs dataset (Dosovitskiy et al. 2015) to compute optical flow in an end-to-end deep network. FlowNet 2.0 (Ilg et al. 2016) uses stacks of networks from FlowNet and performs significantly better, particularly for small motions. Ranjan and Black (2017) propose a Spatial Pyramid Network that employs a small neural network on each level of an image pyramid to compute optical flow. Their method uses a much smaller number of parameters and achieves similar performance as FlowNet (Dosovitskiy et al. 2015) using the same training data. Sun et al. (2018) use image features in a similar spatial pyramid network achieving state-of-the-art results on optical flow benchmarks. Since the above methods are not trained with human motions, they do not perform well on our Human Optical Flow dataset.

2.3 Optical Flow Datasets

Several datasets have been developed to facilitate training and benchmarking of optical flow methods. Middlebury is limited to small motions (Baker et al. 2011), KITTI is focused on rigid scenes and automotive motions (Geiger et al. 2012), while Sintel has a limited number of synthetic scenes (Butler et al. 2012). These datasets are mainly used for evaluation of optical flow methods and are generally too small to support training neural networks.

To learn optical flow using neural networks, more datasets have emerged that contain examples on the order of tens of thousands of frames. The Flying Chairs (Dosovitskiy et al. 2015) dataset contains about 22,000 samples of chairs moving against random backgrounds. Although it is not very realistic or diverse, it provides training data for neural networks (Ranjan and Black 2017; Dosovitskiy et al. 2015) that achieve reasonable results on optical flow benchmarks. Even more recent datasets (Mayer et al. 2016; Gaidon et al. 2016) for optical flow are especially designed for training deep neural networks. Flying Things (Mayer et al. 2016) contains tens of thousands of samples of random 3D objects in motion. The Creative Flow+ Dataset (Shugrina et al. 2019) contains diverse artistic videos in multiple styles. The Monkaa and Driving scene datasets (Mayer et al. 2016) contain frames from animated scenes and virtual driving respectively. Virtual KITTI (Gaidon et al. 2016) uses graphics to generate scenes like those in KITTI and is two orders of magnitude larger. Recent synthetic datasets (Gaidon et al. 2016) show that synthetic data can train networks that generalize to real scenes.

For human bodies, some works (Barbosa et al. 2018; Ghezelghieh et al. 2016) render images with the non-learned artist-defined MakeHuman model (Bastioni et al. 2007) for 3D pose estimation or person re-identification, correspondingly. However, statistical parametric models learned from 3D scans of a big human population, like SMPL (Loper et al. 2015), capture the real distribution of human body shape. The SURREAL dataset (Varol et al. 2017) uses 3D SMPL human meshes rendered on top of color images to train networks for depth estimation, and body part segmentation. While not fully realistic, they show that this data is sufficient to train methods that generalize to real data. We go beyond these works to address the problem of optical flow.

3 The Human Optical Flow Dataset

Our approach generates a realistic dataset of synthetic human motions by simulating them against different realistic backgrounds. We use parametric models (Romero et al. 2017; Loper et al. 2015) to generate synthetic humans with a wide variety of different human shapes. We employ Blender[2] and its Cycles rendering engine to generate realistic synthetic image frames and optical flow. In this way we create the *Human Optical Flow* dataset, that is comprised of two parts. We first create the *Single-Human Optical Flow* (SHOF) dataset (Ranjan et al. 2018) using the body-only SMPL model (Loper et al. 2015) in images containing a single synthetic human. However, image statistics are different for the single- and multi-person case, as multiple people tend to occlude each other in complicated ways. For this reason we then create the *Multi-Human Optical Flow* (MHOF) dataset to better capture this realistic interaction. To make images even more realistic for MHOF, we replace SMPL (Loper et al. 2015) with the SMPL+H (Romero et al. 2017) model that models the body together with articulated fingers, to have richer motion variation. In the rest of this section, we describe the components of our rendering pipeline, shown in Fig. 2. For easy reference, in Table 1 we summarize the data used to generate the SHOF and MHOF datasets, while in Table 2 we summarize the various tools, Blender passes and parameters used for rendering. In the rest of the section, we describe the modules used for generating the data.

3.1 Human Body Generation

3.1.1 Body Model

A parametrized body model is necessary to generate human bodies in a scene. In the SHOF dataset, we use SMPL (Loper et al. 2015) for generating human body shapes. For the MHOF dataset, we use SMPL+H (Romero et al. 2017) that parametrizes the human body together with articulated fingers for increased realism. The models are parameterized by pose and shape parameters to change the body posture and identity, as shown in Fig. 2. They also contain a UV appearance map that allows us to change the skin tone, face features and clothing texture of the resulting virtual humans.

3.1.2 Body Poses

The next step is articulating the human body with different poses, to create moving sequences. To find such poses, we use 3D MoCap datasets (Ionescu et al. 2014; Sigal et al. 2010) (Carnegie-mellon mocap database) that capture 3D MoCap marker positions, glued onto the skin surface of real human subjects. We then employ MoSh (Loper et al. 2014; Mahmood et al. 2019) that fits our body model to these 3D markers by optimizing over parameters of the body model for articulated pose, translation and shape. The pose specifically is a vector of axis-angle parameters, that describes how to rotate each body part around its corresponding skeleton joint.

For the SHOF dataset, we use the Human3.6M dataset (Ionescu et al. 2014), that contains five subjects for training (S1, S5, S6, S7, S8) and two for testing (S9, S11). Each subject performs 15 actions twice, resulting in 1,559,985 frames for training and 550,727 for testing. These sequences are subsampled at a rate of 16×, resulting in 97,499 training and 34,420 testing poses from Human3.6M.

For the MHOF dataset, we use the CMU (Carnegie-mellon mocap database) and HumanEva (Sigal et al. 2010) MoCap datasets to increase motion variation. From CMU MoCap dataset, we use 2605 sequences of 23 high-level action categories. From the HumanEva dataset, we use more than 10 sequences performing actions from 6 different action categories. To reduce redundant poses and allow for larger motions between frames, sequences are subsampled to 12 fps resulting in 321,873 poses. As a result the final MHOF dataset has 254, 211 poses for training, 32,670 for validation and 34,992 for testing.

3.1.3 Hand Poses

Traditionally MoCap systems and datasets (Ionescu et al. 2014; Sigal et al. 2010) (Carnegie-mellon mocap database) record the motion of body joints, and avoid the tedious capture of detailed hand and finger motion. However, in natural settings, people use their body, hands and fingers to communicate social cues and to interact with the physical world. To enable our methods to learn such subtle motions, it should be represented in our training data. Therefore, we use the SMPL+H model (Romero et al. 2017) and augment the body-only MoCap datasets, described above, with finger motion. Instead of using random finger poses that would

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

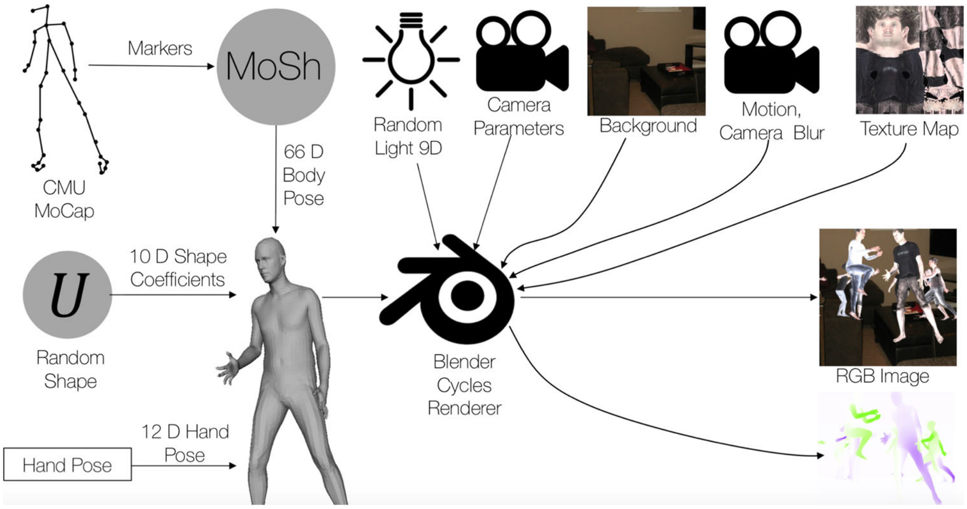

Fig. 2 Pipeline for generating the RGB frames and ground truth optical flow for the *Multi-Human Optical Flow dataset*. The datasets used in this pipeline are listed in Table 1, while the various rendering component are summarized in Table 2

Table 1 Comparison of datasets and most important data preprocessing steps used to generate the SHOF and MHOF datasets

	SHOF	MHOF	Purpose
MoCap data	Human3.6M (Ionescu et al. 2014)	CMU (Gross and Shi 2001), HumanEva (Sigal et al. 2010)	Natural body poses
MoCap → SMPL	MoSh (Loper et al. 2014; Mahmood et al. 2019)	MoSh (Loper et al. 2014; Mahmood et al. 2019)	SMPL parameters from MoCap
Training poses	97,499	254,211	Articulate virtual humans
Validation poses	–	32,670	Articulate virtual humans
Test poses	34,420	34,992	Articulate virtual humans
Hand pose dataset	–	Embodied Hands (Romero et al. 2017)	Natural finger poses
Body shapes	Sample Gaussian distr. (CAESAR) bounded within [− 3, 3] st.dev.	Sample Gaussian distr. (CAESAR) bounded within [− 2.7, 2.7] st.dev.	Body proportions of virtual humans
Textures	CAESAR, non-CAESAR	CAESAR (hands improved), non-CAESAR (hands improved)	Appearance of virtual humans
Background	LSUN (Yu et al. 2015) (indoor) 417,597 images	SUN397 (Xiao et al. 2010) (indoor and outdoor) 30,022 images	Scene background

A short description of the respective part is provided in the last column

generate unrealistic optical flow, we employ the *Embodied Hands* dataset (Romero et al. 2017) and sample continuous finger motion to generate realistic optical flow. We use 43 sequences of hand motion with 37,232 frames recorded at 60 Hz by Romero et al. (2017). Similarly to body MoCap, we subsample hand MoCap to 12 fps to reduce overlapping poses without sacrificing variability.

3.1.4 Body Shapes

Human bodies vary a lot in their proportions, since each person has a unique body shape. To represent this in our dataset, we first learn a gender specific Gaussian distribution of shape parameters, by fitting SMPL to 3D CAESAR scans (Robinette et al. 2002) of both genders. We then sample random body shapes from this distribution to generate a large number of realistic body shapes for rendering. However, naive sampling can result in extreme and unrealistic shape parameters, therefore we bound the shape distribution to avoid unlikely shapes.

For the SHOF dataset, we bound the shape parameters to the range of [− 3, 3] standard deviations for each shape coefficient and draw a new shape for every subsequence of 20 frames to increase variance.

For the MHOF dataset, we account explicitly for collisions and intersections, since intersecting virtual humans would result in generation of inaccurate optical flow. To minimize such cases, we use similar sampling as above with only small differences. We first use shorter subsequences of 10 frames for less frequent inter-human intersections. Furthermore, we bound the shape distribution to the narrower range of [− 2.7, 2.7] standard deviations, since re-targeting

Table 2 Comparison of tools, Blender passes and parameters used to generate the SHOF and MHOF datasets

	SHOF	MHOF	Purpose
Rendering	Cycles	Cycles	Synthetic RGB image rendering
Optical flow	Vector pass (Blender)	Vector pass (Blender)	Optical flow ground truth
Segmentation masks	Material pass (Blender)	Material pass (Blender)	Body part segment. masks (Fig. 3)
Motion blur	Vector pass (Blender)	Vector pass (Blender)	Realistic motion blur artifacts
Imaging noise	Gaussian blur (pixel space) 1px std.dev. for 30% of images	Gaussian blur (pixel space) 1px std.dev. for 30% of images	Realistic image imperfections
Camera translation	Sampled for 30% of frames from Gaussian with 1 cm std.dev.	Sampled for 30% of subsequences from Gaussian with 1 cm std.dev.	Realistic perturbations of the camera (and resulting optical flow)
Camera rotation	Sampled per frame from Gaussian with 0.2° std.dev.	–	Realistic perturbations of the camera (and resulting optical flow)
Illumination	Spherical harmonics (Green 2003)	Spherical harmonics (Green 2003)	Realistic lighting model
Subsequence length	20 frames	10 frames	Number of successive frames with consistent rendering parameters
Mesh collision	–	BVH (Teschner et al. 2004)	Detect (self-)collisions on the triangle level to avoid defect optical flow

The last column provides a short description of the respective method

motion to unlikely body shapes is more prone to mesh self-intersections.

3.1.5 Body Texture

We use the CAESAR dataset (Robinette et al. 2002) to generate a variety of human skin textures. Given SMPL registrations to CAESAR scans, the original per-vertex color in the CAESAR dataset is transferred into the SMPL texture map. Since fiducial markers were placed on the bodies of CAESAR subjects, we remove them from the textures and inpaint them to produce a natural texture. In total, we use 166 CAESAR textures that are of good quality. The main drawback of CAESAR scans is their homogeneity in terms of outfit, since all of the subjects wore grey shorts and the women wore sports bras. In order to increase the clothing variety, we also use textures extracted from our 3D scans (referred as non-CAESAR in the following), to which we register SMPL with 4Cap (Pons-Moll et al. 2015). A total of 772 textures from 7 different subjects with different clothes were captured. We anonymized the textures by replacing the face by the average face in CAESAR, after correcting it to match the skin tone of the texture. Textures are grouped according to the gender, which is randomly selected for each virtual human.

For the SHOF dataset the textures were split in training and testing sets with a 70/30 ratio, while each texture dataset is sampled with a 50% chance. For the MHOF dataset, we introduce more refined splitting with a 80/10/10 ratio for the train, validation and test sets. Moreover, since we introduce also finger motion, we want to favour sampling non-CAESAR textures, due to the bad quality of CAESAR texture maps for the finger region. Thus each texture is sampled with equal probability.

3.1.6 Hand Texture

Hands and fingers are hard to be scanned due to occlusions and measurement limitations. As a result, texture maps are particularly noisy or might even have holes. Since texture is important for optical flow, we augment the body texture maps to improve hand regions. For this we follow a divide and conquer approach. First, we capture hand-only scans with a 3dMD scanner (Romero et al. 2017). Then, we create hand-only textures using the MANO model (Romero et al. 2017), getting 176 high resolution textures from 20 subjects. Finally, we use the hand-only textures to replace the problematic hand regions in the full-body texture maps.

We also need to find the best matching hand-only texture for every body texture. Therefore, we convert all texture maps in HSV space, and compute the mean HSV value for each texture map from standard sampling regions. For full body textures, we sample face regions without facial hair; while for hand-only textures, we sample the center of the outer palm. Then, for each body texture map we find the closest hand-only texture map in HSV space, and shift the values of the latter by the HSV difference, so that the hand skin tone becomes more similar to the facial skin tone. Finally, this improved hand-only texture map is used to replace the pixels in the hand-region of the full body texture map.

3.1.7 (Self-) Collision

The MHOF dataset contains multiple virtual humans moving differently, so there are high chances of collisions and penetrations. This is undesirable because penetrations are physically implausible and unrealistic. Moreover, the generated ground truth optical flow might have artifacts. Therefore, we employ a collision detection method to avoid intersections and penetrations.

Instead of using simple bounding boxes for rough collision detection, we draw inspiration from Tzionas et al. (2016) and perform accurate and efficient collision detection on the triangle level using bounding volume hierarchies (BVH) (Teschner et al. 2004). This level of detailed detection allows for challenging occlusions with small distances between virtual humans, that can commonly be observed for realistic interactions between real humans. This method is useful not only for inter-person collision detection, but also for self-intersections. This is especially useful for our scenarios, as re-targeting body and hand motion to people of different shapes might result in unrealistic self-penetrations. The method is applicable out of the box, with the only exception that we exclude checks of neighboring body parts that are always or frequently in contact, e.g. upper and lower arm, or the two thighs.

3.2 Scene Generation

3.2.1 Background Texture

For the scene background in the SHOF dataset, we use random indoor images from the LSUN dataset (Yu et al. 2015). This provides a good compromise between simplicity and the complex task of generating varied full 3D environments. We use $417,597$ images from the LSUN categories kitchen, living room, bedroom and dining room. These images are placed as billboards, 9 meters from the camera, and are not affected by the spherical harmonics lighting.

In the MHOF dataset, we increase the variability in background appearance, We employ the Sun397 dataset (Xiao et al. 2010) that contains images for 397 highly variable scenes that are both indoor and outdoor, in contrast to LSUN. For quality reasons, we reject all images with resolution smaller than 512×512 px, and also reject images that contain humans using Mask-RCNN (He et al. 2017; Abdulla 2017). As a result, we use $30,222$ images, split in $24,178$ for the training set and $3,022$ for each of the validation and test sets. Further, we increase the distance between the camera and background to 12 meters, to increase the space in which the multiple virtual humans can move without colliding frequently with each other, while still being close enough for visual occlusions.

3.2.2 Scene Illumination

We illuminate the bodies with Spherical Harmonics lighting (Green 2003) that defines basis vectors for light directions. This parameterization is useful for randomizing the scene light by randomly sampling the coefficients with a bias towards natural illumination. The coefficients are uniformly sampled between -0.7 and 0.7, apart from the ambient illumination, which has a minimum value of 0.3 to avoid extremely dark images, and illumination direction, which is strictly negative to favour illumination coming from above.

3.2.3 Increasing Image Realism

In order to increase realism, we introduced three types of image imperfections. First, for 30% of the generated images we introduced camera motion between frames. This motion perturbs the location of the camera with Gaussian noise of 1 cm standard deviation between frames and rotation noise of 0.2 degrees standard deviation per dimension in an Euler angle representation. Second, we added motion blur to the scene using the Vector Blur Node in Blender, and integrated over 2 frames sampled with 64 steps between the beginning and end point of the motion. Finally, we added a Gaussian blur to 30% of the images with a standard deviation of 1 pixel.

3.2.4 Scene Compositing

For animating virtual humans, each MoCap sequence is selected at least once. To increase variability, each sequence is split into subsequences. For the first frame of each subsequence, we sample a body and background texture, lights, blurring and camera motion parameters, and re-position virtual humans on the horizontal plane. We then introduce a random rotation around the z-axis for variability in the motion direction.

For the SHOF dataset, we use subsequences of 20 frames, and at the beginning of each one the single virtual human is re-positioned in the scene such that the pelvis is projected onto the image center.

For the MHOF dataset, we increase the variability with smaller subsequences of 10 frames and introduce more challenging visual occlusions by uniformly sampling the number of virtual humans in the range $[4, 8]$. We sample MoCap sequences S_j with a probability of $p_j = \frac{|S_j|}{\sum_{i=1}^{|S|} |S_i|}$, where $|S_j|$ denotes the number of frames of sequence S_j and $|S|$ the number of sequences. In contrast to the SHOF dataset, for the MHOF dataset the virtual humans are not re-positioned at the center, as they would all collide. Instead, they are placed at random locations on the horizontal plane within camera visibility, making sure there are no collisions with other virtual humans or the background plane during the whole subsequence.

3.3 Ground Truth Generation

3.3.1 Segmentation Masks

Using the material pass of Blender, we store for each frame the ground truth body part segmentation for our models. Although the body part segmentation for both models is similar, SMPL models the palm and fingers as one part, while SMPL+H has a different part segment for each finger bone. Figure 3 shows an example body part segmentation for SMPL+H. These segmentation masks allow us to perform a per body-part evaluation of our optical flow estimation.

3.3.2 Rendering and Ground Truth Optical Flow

For generating images, we use the open source suite Blender and its *vector pass*. The render pass is typically used for producing motion blur, and it produces the motion in image space of every pixel; i.e. the ground truth optical flow. We are mainly interested in the result of this pass, together with the color rendering of the textured bodies.

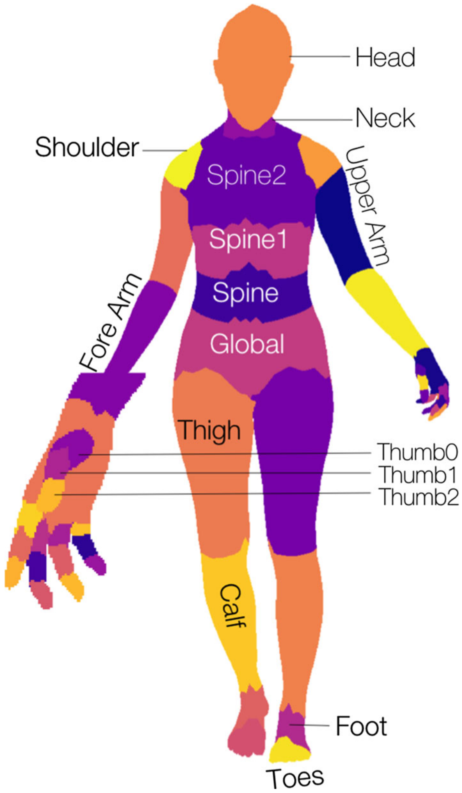

Fig. 3 Body part segmentation for the SMPL+H model. Symmetrical body parts are labeled only once. Finger joints follow the same naming convention as shown for the thumb (best viewed in color)

4 Learning

We train two different network architectures to estimate optical flow on both the SHOF and MHOF dataset. We choose compact models that are based on spatial pyramids, namely SPyNet (Ranjan and Black 2017) and PWC-Net (Sun et al. 2018), shown in Fig. 4. We denote the models trained on the SHOF dataset by SPyNet + SHOF and PWC + SHOF. Similarly, we denote models trained on the MHOF dataset by SPyNet + MHOF and PWC + MHOF.

The spatial pyramid structure employs a convnet at each level of an image pyramid. A pyramid level works on a particular resolution of the image. The top level works on the full resolution and the image features are downsampled as we move to the bottom of the pyramid. Each level learns a convolutional layer d, to perform downsampling of image features. Similarly, a convolution layer u, is learned for decoding optical flow. At each level, the convnet G_k predicts optical flow residuals v_k at that level. These flow residuals get added at each level to produce the full flow, V_K at the finest level of the pyramid.

In SPyNet, each convnet G_k takes a pair of images as inputs along with flow V_{k-1} obtained by resizing the output of the previous level with interpolation. The second frame is however warped using V_{k-1} and the triplet $\{I_k^1, w(I_k^2, V_{k-1}), V_{k-1}\}$ is fed as input to the convnet G_k.

In PWC-Net, a pair of image features, $\{I_k^1, I_k^2\}$ is input at a pyramid level, and the second feature map is warped using using the flow V_{k-1} from the previous level of the pyramid. We then compute the cost-volume $c(I_k^1, w(I_k^2, V_{k-1}))$ over feature maps and pass it to network G_k to compute optical flow V_k at that pyramid level.

We use the pretrained weights as initializations for training both SPyNet and PWC-Net. We train both models end-to-end to minimize the average End Point Error (EPE).

4.1 Hyperparameters

We follow the same training procedure for SPyNet and PWC-Net. The only exception to this is the learning rate, which is determined empirically for each dataset and network from $\{10^{-6}, 10^{-5}, 10^{-4}\}$. For the SHOF we found 10^{-6} to yield best results for SpyNet. Predictions of PWC on the SHOF dataset do not improve for any of these learning rates. For training on MHOF a learning rate of 10^{-6} and 10^{-4} yield best results for SpyNet and PWC-Net, respectively. We use Adam (Kingma and Ba 2014) to optimize our loss with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We use a batch size of 8 and run 400, 000 training iterations. All networks are implemented in the Pytorch framework. Fine-tuning the networks from pretrained weights takes approximately 1 day on SHOF and 2 days on MHOF.

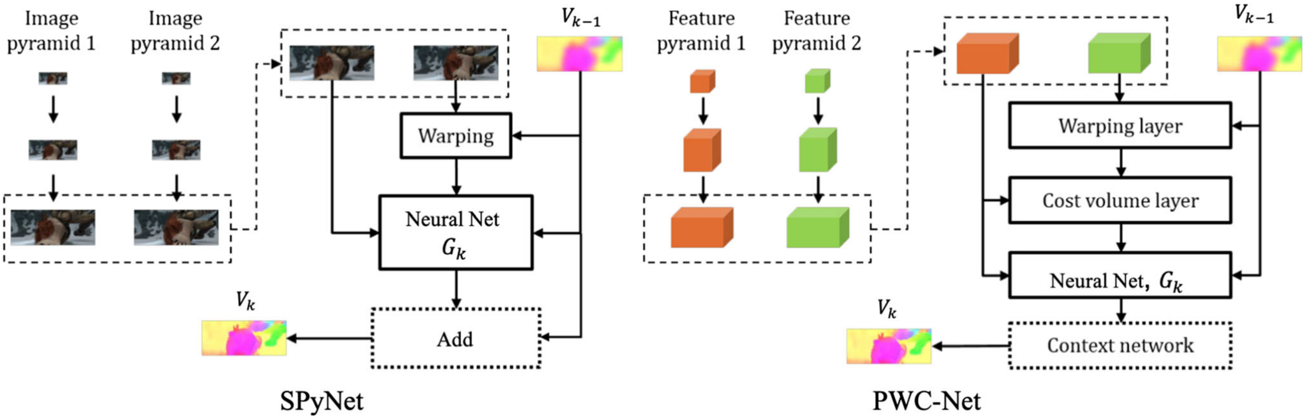

Fig. 4 Spatial Pyramid Network (Ranjan and Black 2017) (left) and PWC-Net (Sun et al. 2018) (right) for optical flow estimation. At each pyramid level, network G_k predicts flow at that level which is used to condition the optical flow at the higher resolution level in the pyramid. Adapted from Sun et al. (2018)

4.2 Data Augmentations

We also augment our data by applying several transformations and adding noise. Although our dataset is quite large, augmentation improves the quality of results on real scenes. In particular, we apply scaling in the range of [0.3, 3], and rotations in [− 17°, 17°]. The dataset is normalized to have zero mean and unit standard deviation using He et al. (2015).

5 Experiments

In this section, we first compare the SHOF , MHOF and other common optical flow datasets. Next, we show that fine-tuning SPyNet on SHOF improves the model, while we observe that fine-tuning PWC-Net on SHOF does not improve the model further. We then fine-tune the same methods on MHOF and evaluate them. We show that both, SPyNet and PWC-Net improve when fine-tuned on MHOF. We show that the methods trained on the MHOF dataset outperform generic flow estimation methods for the pixels corresponding to humans. We show on qualitative results that both, the models trained on SHOF and models trained on MHOF seem to generalize to real word scenes. Finally, we quantitatively evaluate optical flow methods on the MHOF dataset and on a real sequence using motion compensated intensity metric.

5.1 Dataset Details

In comparison with other optical flow datasets, our dataset is larger by an order of magnitude (see Table 3); the SHOF dataset contains 135,153 training frames and 10,867 test frames with optical flow ground truth, while the MHOF dataset has 86,259 training, 13,236 test and 11,817 validation frames. For the single-person dataset we keep the resolution small at 256 × 256 px to facilitate easy deployment for training neural networks. This also speeds up the rendering process in Blender for generating large amounts of data. We show the comparisons of processing time of different models on the SHOF dataset in Table 4. For the MHOF dataset we increase the resolution to 640 × 640 px to be able to reason about optical flow even in small body parts like fingers, using SMPL+H. Our data is extensive, containing a wide variety of human shapes, poses, actions and virtual backgrounds to support deep learning systems.

5.2 Comparison on SHOF

We compare the average End Point Errors (EPEs) of optical flow methods on the SHOF dataset in Table 4, along with the time for evaluation. We show visual comparisons in Fig. 5. Human motion is complex and general optical flow methods fail to capture it. We observe that SPyNet + SHOF outperforms methods that are not trained on SHOF , and SPyNet (Ranjan and Black 2017) in particular. We expect more involved methods like FlowNet2 (Ilg et al. 2016) to have bigger performance gain than SPyNet when trained on SHOF .

We observe that FlowNet (Dosovitskiy et al. 2015) shows poor generalization on our dataset. Since the results of FlowNet (Dosovitskiy et al. 2015) in Table 4 are very close to the zero flow (no motion) baseline, we cross-verify by evaluating FlowNet on a mixture of Flying Chairs (Dosovitskiy et al. 2015) and *Human Optical Flow* and observe that the flow outputs on SHOF is quite random (see Fig. 5). The main reason is that SHOF contains a significant amount of small motions and it is known that FlowNet does not perform very well on small motions. SPyNet + SHOF (Ranjan and Black 2017) however performs quite well and is able to generalize

Table 3 Comparison of the *Human Optical Flow* datasets, namely the *Single-Human Optical Flow* (SHOF) and the *Multi-Human Optical Flow* (MHOF) dataset, with previous optical flow datasets

Dataset	# Train frames	# Test frames	Resolution
MPI Sintel (Butler et al. 2012)	1064	564	1024 × 436
KITTI 2012 (Geiger et al. 2012)	194	195	1226 × 370
KITTI 2015 (Menze and Geiger 2015)	200	200	1242 × 375
Virtual Kitti (Gaidon et al. 2016)	21,260	–	1242 × 375
Flying Chairs (Dosovitskiy et al. 2015)	22,232	640	512 × 384
Flying Things (Mayer et al. 2016)	21,818	4248	960 × 540
Monkaa (Mayer et al. 2016)	8591	–	960 × 540
Driving (Mayer et al. 2016)	4392	–	960 × 540
SHOF (ours)	135,153	10,867	256 × 256
MHOF (ours)	86,259	13,236	640 × 640

Table 4 EPE comparisons and evaluation times of different optical flow methods on the SHOF dataset

Method	AEPE	Time (s)	Learned	Fine-tuned on SHOF
Zero	0.6611	–	–	
FlowNet (Dosovitskiy et al. 2015)	0.5846	0.080	✓	×
PCA Layers (Wulff and Black 2015)	0.3652	10.357	×	×
PWC-Net (Sun et al. 2018)	0.2158	0.024	✓	×
PWC + SHOF	0.2158	0.024	✓	✓
SPyNet (Ranjan and Black 2017)	0.2066	**0.022**	✓	×
Epic Flow (Revaud et al. 2015)	0.1940	1.863	×	×
LDOF (Brox et al. 2009)	0.1881	8.620	×	×
FlowNet2 (Ilg et al. 2016)	0.1895	0.127	✓	×
Flow Fields (Bailer et al. 2015)	0.1709	4.204	×	×
SPyNet + SHOF	**0.1164**	**0.022**	✓	✓

Zero refers to the EPE when zero flow (no motion) is always used for evaluation. Evaluation times are based on the SHOF dataset with 256 × 256 image resolution. We time all GPU based methods using a Tesla V100-16GB GPU
Bold values refers to best performance within the class

to body motions. The results however look noisy in many cases.

Our dataset employs a layered structure where a human is placed against a background. As such layered methods like PCA-layers (Wulff and Black 2015) perform very well on a few images (row 8 in Fig. 5) where they are able to segment a person from the background. However, in most cases, they do not obtain good segmentation into layers.

Previous state-of-the-art methods like LDOF (Brox et al. 2009) and Epic-Flow (Revaud et al. 2015) perform much better than others. They get a good overall shape, and smooth backgrounds. However, their estimation is quite blurred. They tend to miss the sharp edges that are typical of human hands and legs. They are also significantly slower.

In contrast, by fine-tuning on our dataset, the performance of SPyNet + SHOF improves by 40% over SPyNet on the SHOF dataset. We also find that fine-tuning PWC-Net on the SHOF does not improve the model. This could be because SHOF dataset has predominantly small motion which is handled better by SPyNet (Ranjan and Black 2017) architecture. Empirically, we have seen that PWC-Net has state-of-the-art performance on standard benchmarks. This motivates the generation of the MHOF dataset, which includes larger motions and more complex scenes with occlusions.

A qualitative comparison to popular optical flow methods can be seen in Fig. 5. Flow estimations of SPyNet + SHOF can be observed to be sharper than those of methods that are not trained on human motion. This can especially be seen for edges.

5.3 Comparison on MHOF

Training (fine-tuning) on the MHOF dataset improves SPyNet and PWC-Net on average, as can be seen in Table 5. In particular PWC + MHOF outperforms SPyNet+MHOF and also improves over generic state-of-the-art optical flow methods. Large parts of the image are background, whose movements are relatively easy to estimate. However, we are particularly

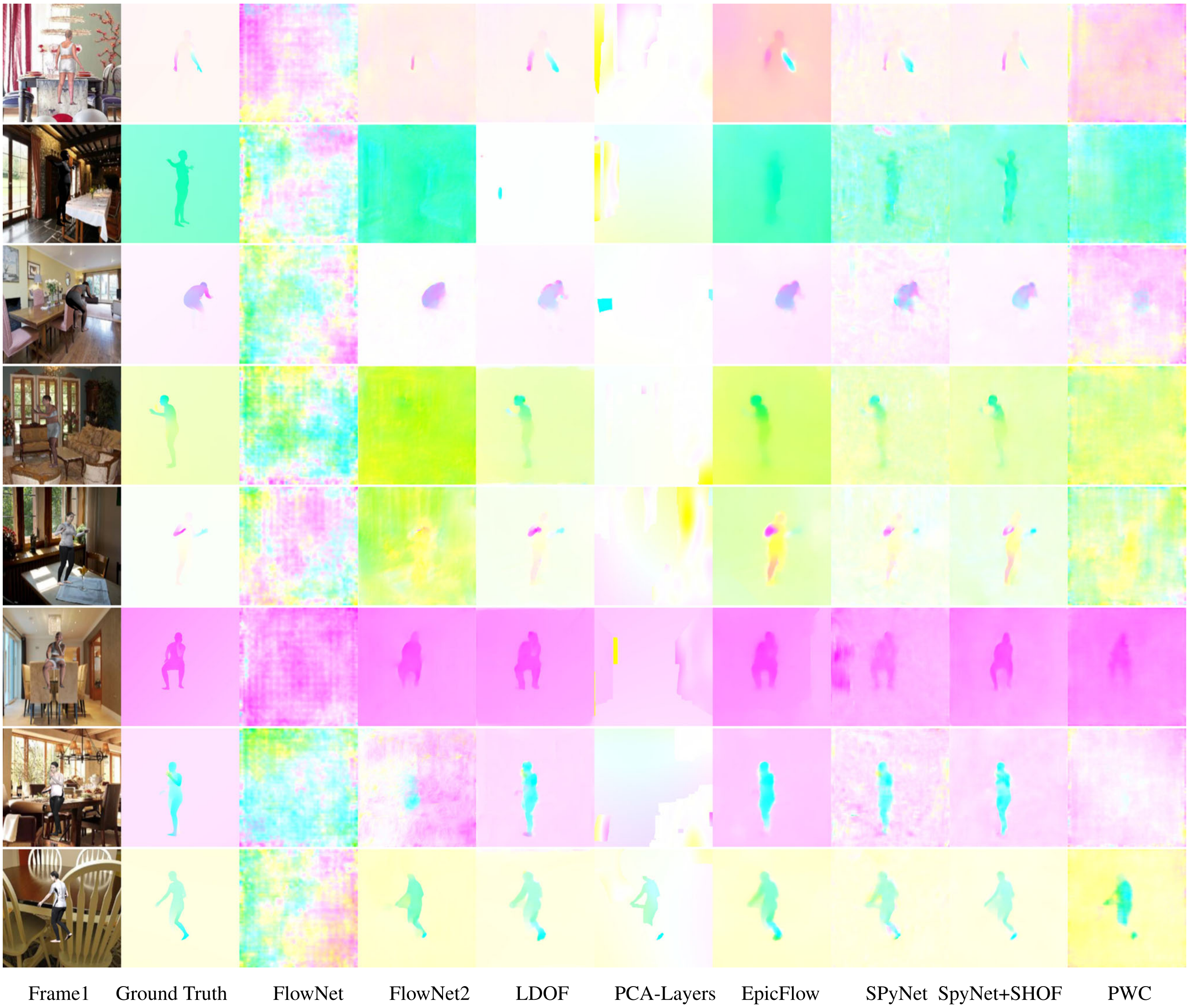

Fig. 5 Visual comparison of optical flow estimates using different methods on the *Single-Human Optical Flow* (SHOF) test set. From left to right, we show Frame 1, Ground Truth flow, results of FlowNet (Dosovitskiy et al. 2015), FlowNet2 (Ilg et al. 2016), LDOF (Brox et al. 2009), PCA-Layers (Wulff and Black 2015), EpicFlow (Revaud et al. 2015), SPyNet (Ranjan and Black 2017), SPyNet + SHOF (ours) and PWC-Net (Sun et al. 2018)

interested in human motions. Therefore, we mask out all errors of background pixels and compute the average EPE only on body pixels (see Table 5). For these pixels, lightweight networks like SpyNet and PWC-Net improve over almost all generic optical flow estimation methods using our dataset (SpyNet + MHOF and PWC + MHOF), including the much larger network FlowNet2. PWC + MHOF is the best performing method (Table 6).

A more fine grained analysis of EPE across body parts is shown in Table 7. We obtain EPE of these body parts using the segmentation shown in Fig. 3. It can be seen that improvements of PWC + MHOF over FlowNet2 are larger for body parts that are at the end of the kinematic tree (i.e. feet, calves, arms and in particular fingers). Differences are less strong for body parts close to the torso. One interpretation of these findings is that movements of the torso are easier to predict, while movements of body parts at the end of the kinematic tree are more complex and thus harder to estimate. In contrast, SPyNet + MHOF outperforms FlowNet2 on body parts close to the torso and does not learn to capture the more complex motions of limbs better than FlowNet2. We expect FlowNet2 + MHOF to perform even better, but we do not include this here due to its long and tedious training process.

Visual comparisons are shown in Fig. 6. In particular, PWC + MHOF predicts flow fields with sharper edges than generic methods or SPyNet + MHOF. Furthermore, the

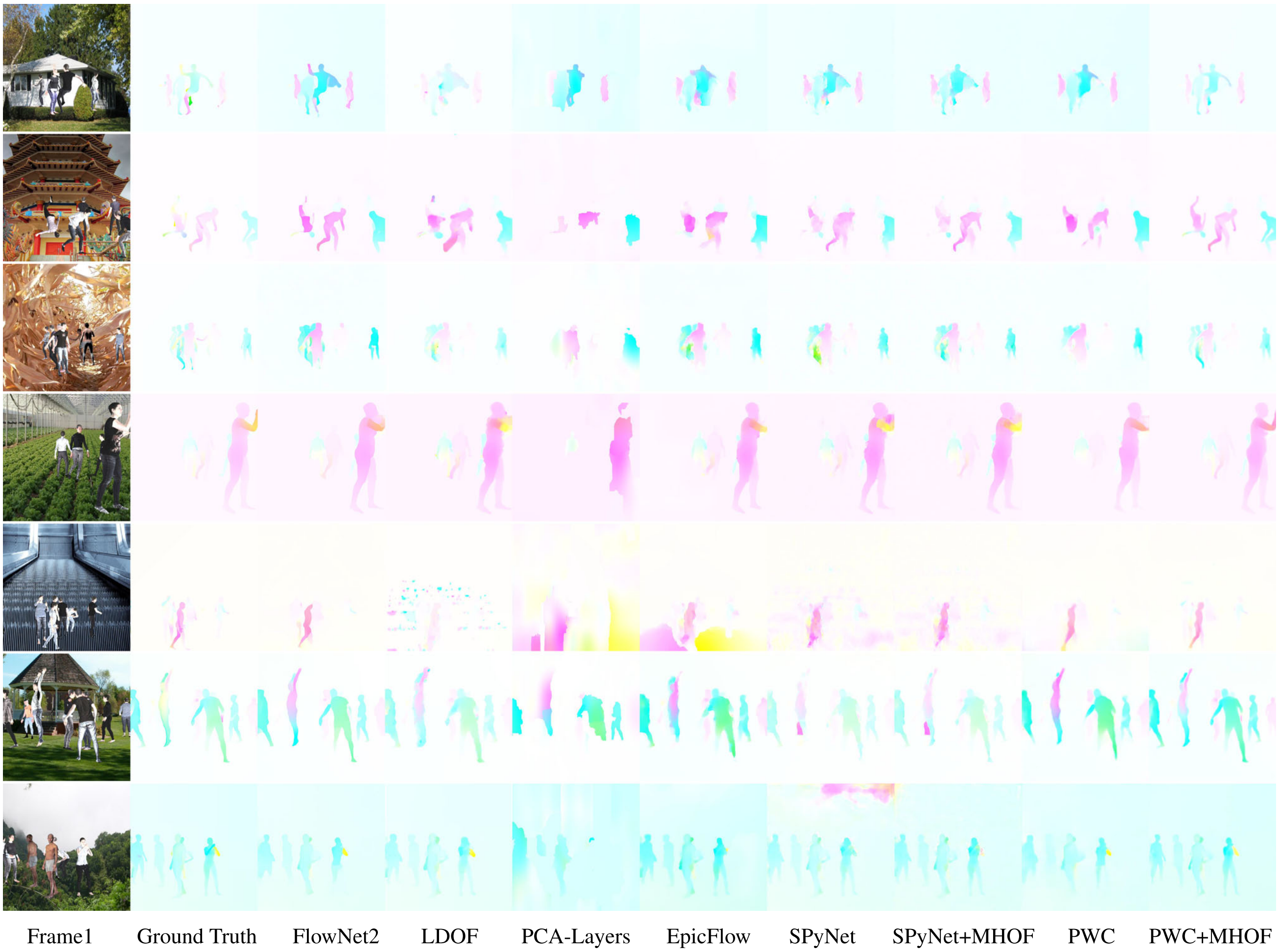

Fig. 6 Visual comparison of optical flow estimates using different methods on the *Multi-Human Optical Flow* (MHOF) test set. From left to right, we show Frame 1, Ground Truth flow, results of FlowNet2 (Ilg et al. 2016), LDOF (Brox et al. 2009), PCA-Layers (Wulff and Black 2015), EpicFlow (Revaud et al. 2015), SPyNet (Ranjan and Black 2017), SPyNet + MHOF (ours), PWC-Net (Sun et al. 2018) and PWC + MHOF (ours)

Table 5 Comparison using End Point Error (EPE) on the *Multi-Human Optical Flow* (MHOF) dataset

Method	Average EPE	Average EPE on *body pixels*	Fine-tuned on MHOF
FlowNet	0.808	2.574	×
PCA Layers	0.556	2.691	×
Epic Flow	0.488	1.982	×
SPyNet	0.429	1.977	×
SPyNet + MHOF	0.391	1.803	✓
PWC-Net	0.369	2.056	×
LDOF	0.360	1.719	×
FlowNet2	0.310	1.863	×
PWC + MHOF	**0.301**	**1.621**	✓

We show the average EPE and body-only EPE. For the latter, the EPE is computed only over segments of the image depicting a human body. Best results are shown in boldface. A comparison of body-part specific EPE can be found in Table 7

Table 6 Comparison using Motion Compensated Intensity (MCI) on the *Multi-Human Optical Flow* (MHOF) dataset and a real video sequence

Method	Average MCI MHOF	Average MCI real
FlowNet	287.328	401.779
PCA layers	201.594	423.332
Epic flow	129.252	234.037
SPyNet	142.108	302.753
SPyNet + MHOF	143.029	297.142
PWC-Net	157.088	344.202
LDOF	**71.449**	**158.281**
FlowNet2	145.732	303.799
PWC + MHOF	152.314	351.567

Example images for the real video sequence can be seen in Fig. 9
Bold values refers to best performance within the class

Table 7 Comparison using End Point Error (EPE) on the *Multi-Human Optical Flow* (MHOF) dataset

Parts	Epic flow	LDOF	FlowNet2	FlowNet	PCA layers	PWC-Net	PWC + MHOF	SPyNet	SPyNet + MHOF
Average (whole image)	0.488	0.360	0.310	0.808	0.556	0.369	**0.301**	0.429	0.391
Average (body pixels)	1.982	1.719	1.863	2.574	2.691	2.056	**1.621**	1.977	1.803
global	1.269	1.257	1.337	2.005	1.920	1.389	**1.163**	1.356	1.236
head	1.806	**1.328**	1.626	2.681	2.808	1.881	1.445	1.708	1.519
leftCalf	2.116	1.802	1.787	2.420	2.711	2.109	**1.476**	1.991	1.796
leftFoot	3.089	2.346	2.476	2.987	3.393	3.002	**2.142**	2.701	2.566
leftForeArm	3.972	3.231	3.536	4.380	4.778	3.926	**3.136**	3.945	3.605
leftHand	5.777	4.422	4.823	5.928	6.531	5.634	**4.385**	5.547	5.040
leftShoulder	1.513	**1.429**	1.646	2.331	2.336	1.732	1.471	1.560	1.462
leftThigh	1.424	1.338	1.466	2.102	2.150	1.565	**1.230**	1.517	1.362
leftToes	3.147	2.573	2.755	3.065	3.307	3.100	**2.524**	2.830	2.784
leftUpperArm	2.215	**1.947**	2.288	3.005	3.139	2.376	1.955	2.307	2.076
lIndex0	6.199	4.900	5.334	6.254	6.785	6.124	**4.861**	5.925	5.472
lIndex1	6.367	**5.159**	5.672	6.340	6.829	6.303	5.212	6.087	5.727
lIndex2	6.315	**5.253**	5.878	6.203	6.670	6.270	5.433	6.028	5.784
lMiddle0	6.338	4.983	5.331	6.364	6.910	6.211	**4.837**	6.012	5.544
lMiddle1	6.498	5.239	5.632	6.435	6.927	6.383	**5.176**	6.143	5.767
lMiddle2	6.266	**5.212**	5.756	6.130	6.592	6.182	5.303	5.934	5.679
lPinky0	6.048	**4.792**	5.302	6.035	6.603	5.940	4.873	5.738	5.307
lPinky1	6.106	**4.922**	5.489	6.038	6.574	6.014	5.064	5.765	5.418
lPinky2	5.780	**4.856**	5.419	5.655	6.170	5.702	4.956	5.474	5.231
lRing0	6.388	4.973	5.281	6.413	7.010	6.218	**4.834**	6.064	5.552
lRing1	6.313	5.083	5.391	6.256	6.801	6.168	**4.949**	5.966	5.558
lRing2	6.047	**5.035**	5.515	5.924	6.409	5.942	5.067	5.710	5.441
lThumb0	5.415	**4.318**	4.673	5.473	6.072	5.316	4.329	5.212	4.809
lThumb1	5.636	**4.527**	5.065	5.698	6.232	5.612	4.685	5.449	5.065
lThumb2	5.825	**4.749**	5.388	5.820	6.323	5.802	5.005	5.629	5.314
neck	1.336	**1.195**	1.371	2.151	2.245	1.440	1.227	1.399	1.250
rightCalf	2.243	1.892	1.864	2.530	2.851	2.223	**1.548**	2.081	1.907
rightFoot	3.270	2.454	2.610	3.149	3.599	3.171	**2.276**	2.894	2.732
rightForeArm	3.990	3.242	3.554	4.381	4.759	3.928	**3.190**	4.029	3.641
rightHand	5.735	4.348	4.787	5.837	6.447	5.550	**4.339**	5.582	4.978
rightShoulder	1.547	**1.431**	1.670	2.390	2.340	1.735	1.477	1.573	1.462
rightThigh	1.477	1.374	1.512	2.158	2.226	1.624	**1.263**	1.556	1.407

Table 7 continued

Parts	Epic flow	LDOF	FlowNet2	FlowNet	PCA layers	PWC-Net	PWC + MHOF	SPyNet	SPyNet + MHOF
rightToes	3.395	2.707	2.918	3.293	3.566	3.346	**2.699**	3.064	2.999
rightUpperArm	2.267	**1.974**	2.294	3.033	3.148	2.400	2.007	2.002	2.113
rIndex0	6.264	**4.875**	5.324	6.255	6.800	6.150	4.886	6.003	5.486
rIndex1	6.541	**5.210**	5.755	6.449	6.951	6.457	5.329	6.237	5.835
rIndex2	6.465	**5.320**	5.968	6.294	6.776	6.404	5.533	6.149	5.879
rMiddle0	6.509	5.056	5.454	6.470	7.014	6.354	**4.967**	6.211	5.662
rMiddle1	6.680	5.341	5.777	6.562	7.058	6.537	**5.325**	6.325	5.895
rMiddle2	6.394	**5.261**	5.838	6.209	6.713	6.274	5.366	6.038	5.739
rPinky0	5.983	**4.750**	5.372	5.952	6.504	5.855	4.845	5.741	5.262
rPinky1	6.076	**4.905**	5.566	5.979	6.533	5.943	5.025	5.809	5.402
rPinky2	5.789	**4.813**	5.403	5.645	6.220	5.662	4.903	5.532	5.232
rRing0	6.397	4.948	5.350	6.383	6.938	6.215	**4.856**	6.126	5.565
rRing1	6.395	5.108	5.465	6.290	6.841	6.212	**5.019**	6.066	5.615
rRing2	6.222	**5.129**	5.644	6.052	6.610	6.063	5.160	5.889	5.571
rThumb0	5.417	**4.304**	4.748	5.470	6.057	5.301	4.360	5.247	4.819
rThumb1	5.605	**4.465**	4.945	5.643	6.210	5.514	4.607	5.434	5.032
rThumb2	5.835	**4.748**	5.262	5.789	6.328	5.749	4.938	5.639	5.306
spine	1.233	1.271	1.325	1.941	1.856	1.360	**1.168**	1.322	1.221
spine1	1.330	1.369	1.421	2.028	1.957	1.460	**1.268**	1.417	1.322
spine2	1.329	1.308	1.439	2.089	2.049	1.480	**1.276**	1.387	1.309

We show the average EPE and body part specific EPE, where part labels follow Fig. 3. The first two rows are repeated from Table 5
Bold values refers to best performance within the class

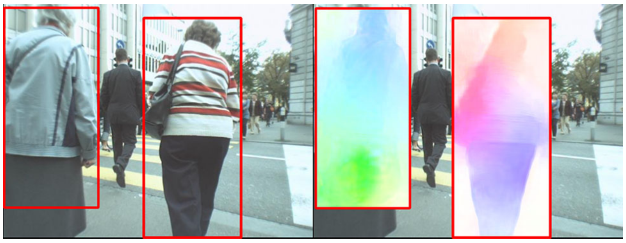

Fig. 7 We use the DPM (Felzenszwalb et al. 2010) person detector to crop out people from real-world scenes (left) and use SPyNet + SHOF to compute optical flow on the cropped section (right)

qualitative results suggest that PWC + MHOF is better at distinguishing the motion of people, as people can be better separated on the flow visualizations of PWC + MHOF (Fig. 6, row 3). Last, it can be seen that fine details, like the motion of distant humans or small body parts, are better estimated by PWC + MHOF.

The above observations are strong indications that our *Human Optical Flow* datasets (SHOF and MHOF) can be beneficial for the performance on human motion for other optical flow networks as well.

5.4 Real Scenes

We show a visual comparison of results on real-world scenes of people in motion. For visual comparisons of models trained on the SHOF dataset we collect these scenes by cropping people from real world videos as shown in Fig. 7. We use DPM (Felzenszwalb et al. 2010) for detecting people and compute bounding box regions in two frames using the ground truth of the MOT16 dataset (Milan et al. 2016). The results for the SHOF dataset are shown in Fig. 8. A comparison of methods on real images with multiple people can be seen in Fig. 9.

The performance of PCA-Layers (Wulff and Black 2015) is highly dependent on its ability to segment. Hence, we see only a few cases where it looks visually correct. SPyNet (Ranjan and Black 2017) gets the overall shape but the results look noisy in certain image parts. While LDOF (Brox et al. 2009), EpicFlow (Revaud et al. 2015) and FlowFields (Bailer et al. 2015) generally perform well, they often find it difficult to resolve the legs, hands and head of the person. The results from models trained on our *Human Optical Flow* dataset look appealing especially while resolving the overall human shape, and various parts like legs, hands and the human head. Models trained on the *Human Optical Flow* dataset perform well under occlusion (Figs. 8, 9). Many examples including severe occlusion can be seen in Fig. 9. Besides that, Fig. 9 shows that the models trained on MHOF are able to distinguish motions of multiple people and predict sharp edges of humans.

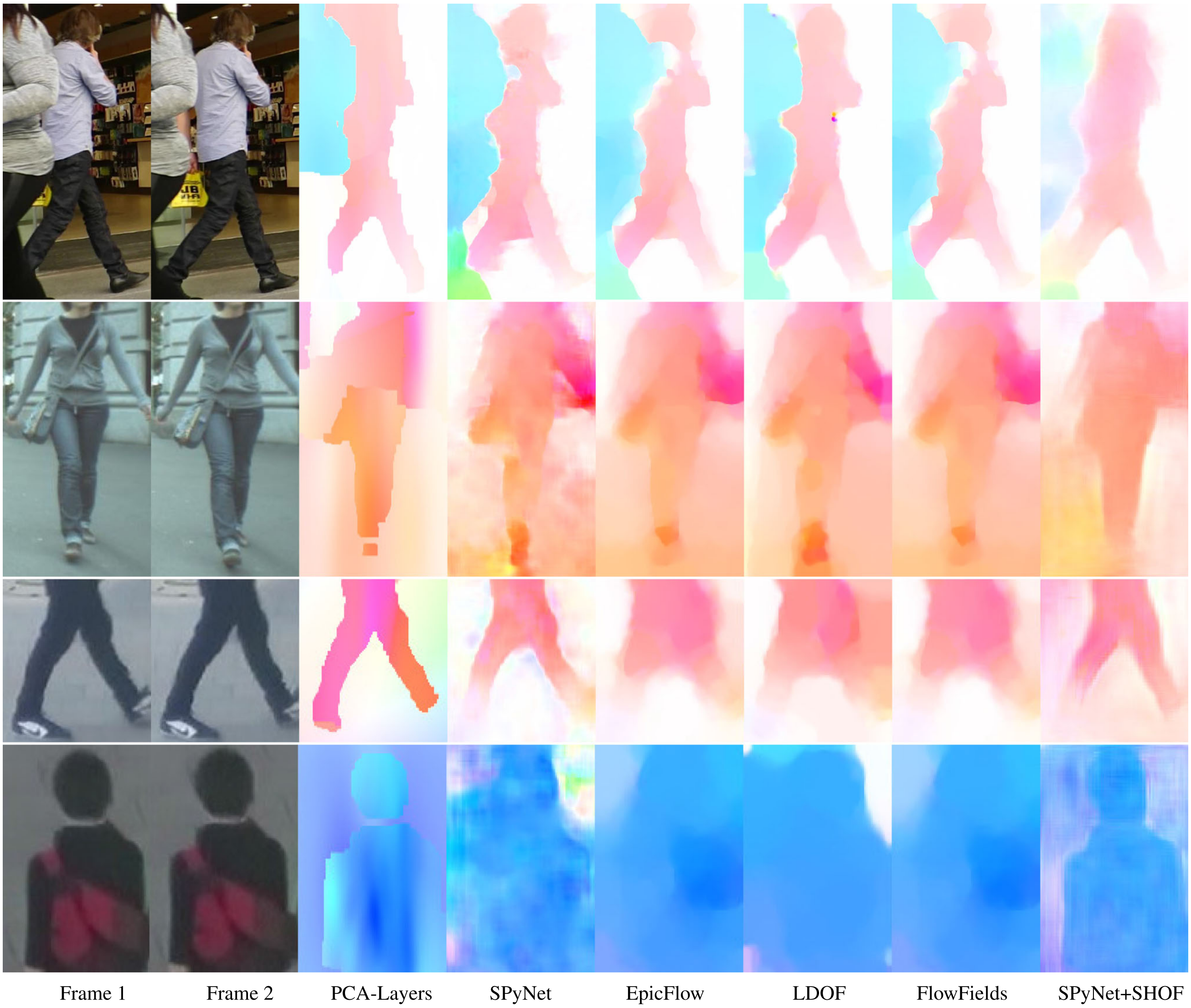

Fig. 8 *Single-Human Optical Flow*visuals on real images using different methods. From left to right, we show Frame 1, Frame 2, results of PCA-Layers (Wulff and Black 2015), and SPyNet (Ranjan and Black 2017), EpicFlow (Revaud et al. 2015), LDOF (Brox et al. 2009), FlowFields (Bailer et al. 2015) and SPyNet + SHOF (ours)

A quantitative evaluation on real data with humans is not possible, as no such dataset with ground truth optical flow annotation exists. To determine generalization of the models to real data, despite the lack of ground truth annotation, we can use the Motion Compensated Intensity (MCI) as an error metric. Given the image sequence I^1, I^2 and predicted flow V, the MCI error is given by

$$\mathrm{MCI}(I^1, I^2, V) = ||I^1 - w(I^2, V)||^2, \qquad (1)$$

where w warps the image I^2 according to flow V. This metric certainly has limitations. The motion compensated intensity assumes Lambertian conditions i.e. intensity of a point remains constant over time. MCI error does not account for occlusions. Furthermore, MCI does not account for smooth flow fields over texture-less surfaces. Despite these shortcoming of MCI, we report these numbers to show that our models generalize to real data. However, it should be noted that EPE is a more precise metric to evaluate optical flow estimation.

To test whether MCI correlates with EPEs in Table 5, we compute MCI on the MHOF dataset. The results can be seen in Table 6. We observe that, methods like FlowNet and PCA-Layers which have poor performance on the EPE metric have higher MCI error. For methods with lower EPE, the MCI errors do not exactly correspond to the respective EPEs. This

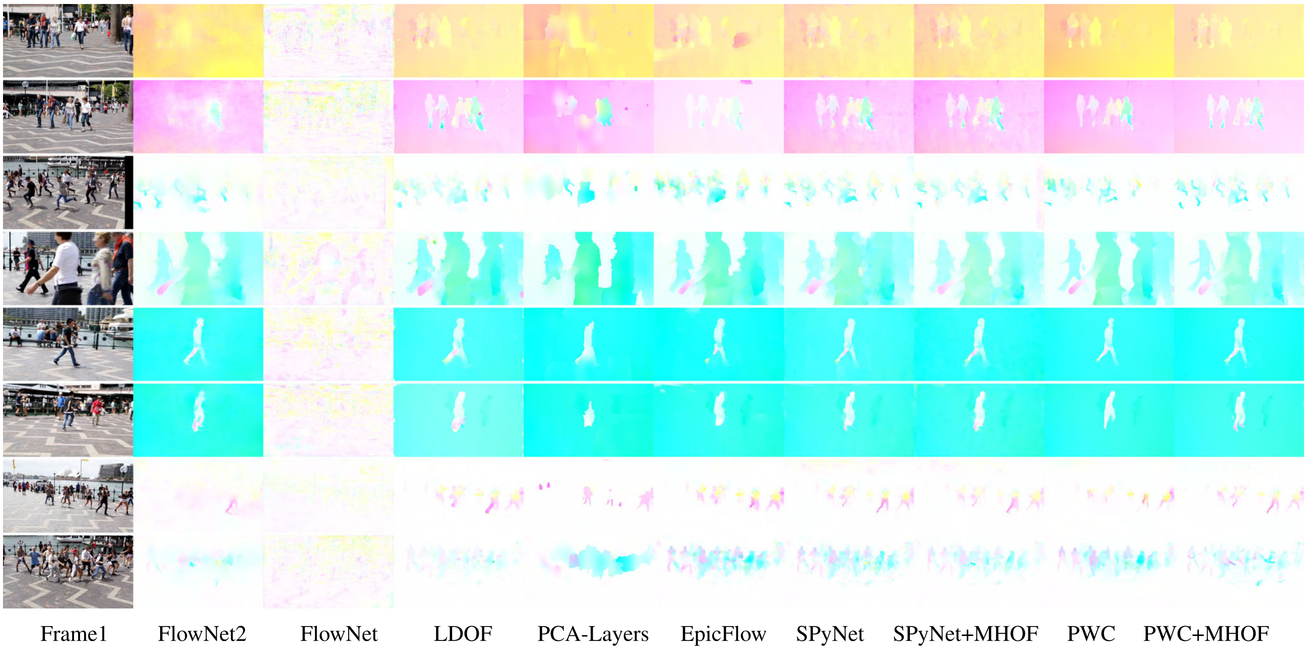

Fig. 9 *Multi-Human Optical Flow* visuals on real images. From left to right, we show Frame 1, results of FlowNet2 (Ilg et al. 2016), FlowNet (Dosovitskiy et al. 2015), LDOF (Brox et al. 2009), PCA-Layers (Wulff and Black 2015), EpicFlow (Revaud et al. 2015), SPyNet (Ranjan and Black 2017), SPyNet + MHOF(ours), PWC-Net (Sun et al. 2018) and PWC + MHOF (ours)

is due to the limitations of the MCI metric, as described above. Finally, we compute MCI on a real video sequence from Youtube.[3] The MCI errors are shown in Table 6.

6 Conclusion and Future Work

In summary, we created an extensive *Human Optical Flow* dataset containing images of realistic human shapes in motion together with ground truth optical flow. The dataset is comprised of two parts, the *Single-Human Optical Flow* (SHOF) and the *Multi-Human Optical Flow* (MHOF) dataset. We then train two compact network architectures based on spatial pyramids, namely SpyNet and PWC-Net. The realism and extent of our dataset, together with an end-to-end training scheme, allows these networks to outperform previous state-of-the-art optical flow methods on our new human-specific dataset. This indicates that our dataset can be beneficial for other optical flow network architectures as well. Furthermore, our qualitative results suggest that the networks trained on the *Human Optical Flow* generalize well to real world scenes with humans. This is evidenced by results on a real sequence using the MCI metric. The trained models are compact and run in real time making them highly suitable for phones and embedded devices.

The dataset and our focus on human optical flow opens up a number of research directions in human motion understanding and optical flow computation. We would like to extend our dataset by modeling more diverse clothing and outdoor scenarios. A direction of potentially high impact for this work is to integrate it in end-to-end systems for action recognition, which typically take precomputed optical flow as input. The real-time nature of the method could support motion-based interfaces, potentially even on devices like cell phones with limited computing power. The dataset, dataset generation code, pretrained models, and training code are available, enabling researchers to use them for problems involving human motion.

Acknowledgements Open access funding provided by Projekt DEAL. We thank Yiyi Liao for helping us with optical flow evaluation. We thank Sergi Pujades for helping us with collision detection of meshes. We thank Cristian Sminchisescu for the Human3.6M MoCap marker data.

[3] https://www.youtube.com/watch?v=2DiQUX11YaY.

References

Abdulla, W. (2017). Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN.

Bailer, C., Taetz, B., & Stricker, D. (2015). Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation. In *Proceedings of the IEEE international conference on computer vision* (pp. 4015–4023).

Baker, S., Scharstein, D., Lewis, J. P., Roth, S., Black, M. J., & Szeliski, R. (2011). A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, *92*(1), 1–31.

Barbosa, I. B., Cristani, M., Caputo, B., Rognhaugen, A., & Theoharis, T. (2018). Looking beyond appearances: Synthetic training data for deep cnns in re-identification. *Computer Vision and Image Understanding*, *167*, 50–62.

Bastioni, M., Flerackers, M., & Capco J. (2007). Makehuman: Open source tool for making 3d characters.

Black, M. J., Yacoob, Y., Jepson, A. D., & Fleet, D. J. (1997). Learning parameterized models of image motion. In *IEEE conference on computer vision and pattern recognition, CVPR-97*, Puerto Rico (pp. 561–567).

Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., & Black, M. J. (2016). Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer vision—ECCV 2016*, Lecture Notes in Computer Science. Springer.

Brox, T., Bregler, C., & Malik, J. (2009). Large displacement optical flow. In *IEEE conference on computer vision and pattern recognition, 2009. CVPR 2009* (pp. 41–48). IEEE.

Butler, D. J., Wulff, J., Stanley, G. B., & Black, M. J. (2012). A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), *European conference on computer vision (ECCV)*, Part IV, LNCS 7577 (pp. 611–625). Springer.

Charles, J., Pfister, T., Magee, D. R., Hogg, D. C., & Zisserman, A. (2016). Personalizing human video pose estimation. In *CVPR* (pp. 3063–3072). IEEE Computer Society.

Davis, J. W. (2001). Hierarchical motion history images for recognizing human motion. In *Detection and recognition of events in video* (pp. 39–46).

Dong, X., Yu, S.-I., Weng, X., Wei, S.-E., Yang, Y., & Sheikh, Y. (2018). Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Dosovitskiy, A., Fischery, P., Ilg, E., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., & Brox, T., et al. (2015). Flownet: Learning optical flow with convolutional networks. In *2015 IEEE international conference on computer vision (ICCV)* (pp. 2758–2766). IEEE.

Fablet, R., & Black, M. J. (2002). Automatic detection and tracking of human motion with a view-based representation. In *European conference on computer vision, ECCV 2002*, volume 1 of *LNCS 2353* (pp. 476–491). Springer.

Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In *CVPR* (pp. 1933–1941). IEEE Computer Society.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *TPAMI*, *32*(9), 1627–1645.

Fragkiadaki, K., Hu, H., & Shi, J. (2013). Pose from flow and flow from pose. In *2013 IEEE conference on computer vision and pattern recognition* (pp. 2059–2066).

Freeman, W. T., Pasztor, E. C., & Carmichael, O. T. (2000). Learning low-level vision. *International Journal of Computer Vision*, *40*(1), 25–47.

Gaidon, A., Harchaoui, Z., & Schmid, C. (2014). Activity representation with motion hierarchies. *International Journal of Computer Vision*, *107*(3), 219–238.

Gaidon, A., Wang, Q., Cabon, Y., & Vig, E. (2016). Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*.

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Conference on computer vision and pattern recognition (CVPR)*.

Geman, D., & Geman, S. (2016). Opinion: Science in the age of selfies. *Proceedings of the National Academy of Sciences*, *113*(34), 9384–9387.

Ghezelghieh, M. F., Kasturi, R., & Sarkar, S. (2016). Learning camera viewpoint using CNN to improve 3D body pose estimation. In *2016 fourth international conference on 3D vision (3DV)* (pp. 685–693). IEEE.

Green, R. (2003). Spherical harmonic lighting: The gritty details. In *Archives of the game developers conference*.

Gross, R., & Shi, J. (2001). The cmu motion of body (mobo) database.

Güney, F., & Geiger, A. (2016). Deep discrete flow. In *Asian conference on computer vision (ACCV)*.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *2017 IEEE international conference on computer vision (ICCV)* (pp. 2980–2988). IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. arXiv:1512.03385.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., & Brox, T. (2016). Flownet 2.0: Evolution of optical flow estimation with deep networks. arXiv:1612.01925.

Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2014). Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1325–1339.

Jhuang, H., Gall, J., Zuffi, S., Schmid, C., & Black, M. J. (2013). Towards understanding action recognition. In *IEEE international conference on computer vision (ICCV)*, Sydney, Australia, December 2013 (pp. 3192–3199). IEEE.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*(2), 201–211.

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv:1412.6980.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). Hmdb: A large video database for human motion recognition. In *2011 IEEE international conference on computer vision (ICCV)* (pp. 2556–2563). IEEE.

Loper, M., Mahmood, N., & Black, M. J. (2014). Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, *33*(6), 220.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2015). SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, *34*(6), 248:1–248:16.

Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., Black, M. J. (2019). AMASS: Archive of motion capture as surface shapes. arXiv:1904.03278.

Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., & Brox, T. (2016). A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE international conference on computer vision and pattern recognition (CVPR)*. arXiv:1512.02134.

Menze, M., & Geiger, A. (2015). Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3061–3070).

Milan, A., Leal-Taixé, L., Reid, I. D., Roth, S., & Schindler, K. (2016). MOT16: A benchmark for multi-object tracking. arXiv:1603.00831.

Pfister, T., Charles, J., & Zisserman, A. (2015). Flowing convnets for human pose estimation in videos. In *ICCV* (pp. 1913–1921). IEEE Computer Society.

Pons-Moll, G., Romero, J., Mahmood, N., & Black, M. J. (2015). Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, *34*(4), 120:1–120:14.

Ranjan, A., & Black, M. J. (2017). Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Ranjan, A., Romero, J., & Black, M. J. (2018). Learning human optical flow. In *29th British machine vision conference*.

Revaud, J., Weinzaepfel, P., Harchaoui, Z., & Schmid, C. (2015). EpicFlow: Edge-preserving interpolation of correspondences for optical flow. In *Computer vision and pattern recognition*.

Robinette, K. M., Blackwell, S., Daanen, H., Boehmer, M., & Fleming, S. (2002). Civilian American and European surface anthropometry resource (caesar), final report, volume 1. summary. Technical report, DTIC Document.

Romero, J., Loper, M., & Black, M. J. (2015). FlowCap: 2D human pose from optical flow. In *Proceedings of 37th German conference on pattern recognition (GCPR) pattern recognition*, volume LNCS 9358 (pp. 412–423). Springer.

Romero, J., Tzionas, D., & Black, M. J. (2017). Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, *36*(6), 245.

Shugrina, M., Liang, Z., Kar, A., Li, J., Singh, A., Singh, K., & Fidler, S. (2019). Creative flow+ dataset. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Sigal, L., Balan, A., & Black, M. J. (2010). HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, *87*(1), 4–27.

Soomro, K., Zamir, A. R., & Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402.

Sun, D., Roth, S., Lewis, J. P., & Black, M. J. (2008). Learning optical flow. In *ECCV* (pp. 83–97).

Sun, D., Yang, X., Liu, M. Y., & Kautz, J. (2018). PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*.

Teschner, M., Kimmerle, S., Heidelberger, B., Zachmann, G., Raghupathi, L., Fuhrmann, A., Cani, M.-P., Faure, F., Magnenat-Thalmann, N., Strasser, W., & Volino, P. (2004). Collision detection for deformable objects. In *Eurographics* (pp. 119–139).

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2016). Deep end2end voxel2voxel prediction. In *The 3rd workshop on deep learning in computer vision*.

Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M., & Gall, J. (2016). Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision (IJCV)*, *118*(2), 172–193.

Varol, G., Romero, J., & Martin, X., Mahmood, N., Black, M. J., Laptev, I., & Schmid, C. (2017). Learning from synthetic humans. In *CVPR*.

Wulff, J., & Black, M. J. (2015). Efficient sparse-to-dense optical flow estimation using a learned basis and layers. In *2015 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 120–130). IEEE.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3485–3492). IEEE.

Yu, F., Zhang, Y., Song, S., Seff, A., & Xiao, J. (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv:1506.03365.

Zuffi, S., Romero, J., Schmid, C., & Black, M. J. (2013). Estimating human pose with flowing puppets. In *IEEE international conference on computer vision (ICCV)* (pp. 3312–3319).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.