Bootstrapping Human Optical Flow and Pose

Aritro Roy Arko
aritroroyarko30@gmail.com
James J. Little
little@cs.ubc.ca
Kwang Moo Yi
kmyi@cs.ubc.ca

Computer Vision Lab
The University of British Columbia
British Columbia, Canada

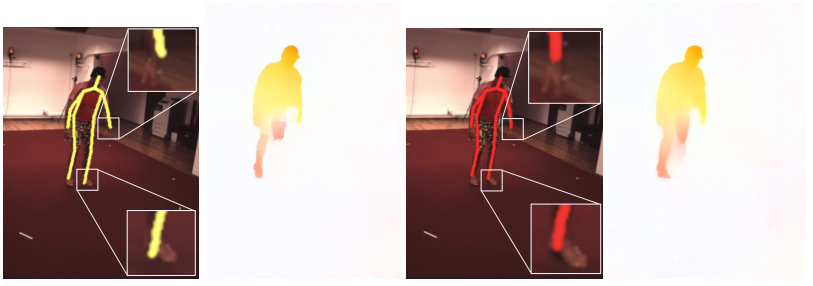
Abstract

We propose a bootstrapping framework to enhance human optical flow and pose. We show that, for videos involving humans in scenes, we can improve both the optical flow and the pose estimation quality of humans by considering the two tasks at the same time. We enhance optical flow estimates by fine-tuning them to fit the human pose estimates and vice versa. In more detail, we optimize the pose and optical flow networks to, at inference time, agree with each other. We show that this results in state-of-the-art results on the Human 3.6M and 3D Poses in the Wild datasets, as well as a human-related subset of the Sintel dataset, both in terms of pose estimation accuracy and the optical flow accuracy at human joint locations. Code available at https://github.com/ubc-vision/bootstrapping-human-optical-flow-and-pose

1 Introduction

Estimating the pose and the motion of humans plays an important role in various tasks in Computer Vision, including human activity recognition [38], pedestrian analysis [9], and pose-based medical diagnosis [33]. Naturally, various methods have been proposed to estimate human pose accurately [17, 19] as well to analyze motion (optical flow) in scenes [35]. While these methods have been highly successful in the task of human pose estimation and optical flow estimation, respectively, they focus either on solely on the human pose [17, 19], or focus on generic optical flow [35].

This leaves room for improvement. Regarding estimating motions, while unsurprising, indirect evidence of potential for improvement arises from work focusing on human optical flow [31, 32]. Generic optical flow methods, such as Spynet [30], perform better for estimating optical flow of humans when fine-tuned on human-centric scenes. This demonstrates that when the task revolves around humans, the optical flow method should also focus on humans. On the other hand, regarding human pose, an important but overlooked assumption in recent works is temporal consistency. While [10] utilized multiple frames to take advantage of temporal consistency, this is, in fact, left for the Neural Network to implicitly figure out and embed into the framework while training. However, modern generic optical flow



(a) Pose from [19] (b) Flow from [35] (c) Flow enhanced pose (d) Pose enhanced flow

Figure 1: **Teaser –** Given an input image, we use off-the-shelf methods [19], [35] to estimate the pose and the flow of the scene, which we then improve them *without* any retraining.

methods perform surprisingly well [35], so learning this temporal relationship implicitly is a difficulty one does not have to go through—we already have the tools.

In our work, we propose to make use of the tools that already exist—human pose estimators and optical flow networks—and enhance their performance by marrying the two. Our idea originates from the fact that, should human optical flow and pose be estimated properly, they should coincide—the movement of the joints, when projected in 2D, should follow the optical flow estimates at these locations. We thus create an iterative flow-pose-flow optimization framework for inference, where, assuming both estimators are not completely failing we enhance the optical flow to match the pose estimates, which we then optimize the pose to match the flow, and finally optimize the flow once more to match the pose; See Figure 1 for an example. While this process can be repeated multiple times for further enhancement, with the state-of-the-art methods that are already performing well, we find that this three-round setup is enough for accurate estimates.

In more detail, for the optical flow network we utilize the Recurrent All Pairs Field Transform (RAFT) [35], and for the human pose estimator the Mesh Transformer (METRO) [19]. We then, given a video sequence of a human, fine-tune the RAFT network so that the flow estimates match the METRO pose estimates at the pixels that correspond to the projected 2D locations of the human skeleton, which would then roughly represent how the human moved between two frames. In other words, we create the 2D stick-man representation of the human considering all the bone pixels, and then ensure that the optical flow values at the bone pixels are close to the difference in bone pixel locations in consecutive frames. With enhanced flow, we obtain enhanced poses by optimizing the 3D poses directly to match the flow, while considering also the temporal smoothness of the estimates, similar to [1] and consistent bone length. Finally, we fine-tune the RAFT network once more with the enhanced poses. While we initially utilize RAFT and METRO, these can be trivially replaced with any other human pose estimator and optical flow network as we show in our experiments. We note that, during this entire process, no re-training is required and we are strictly fine-tuning to the *test data without any label*. In other words, we are solving our problem in a setup similar to *transductive* learning.

We validate the efficacy of our method on three datasets: Human 3.6M [12] and 3D Poses in the Wild (3DPW) [41] datasets, both of which are human pose datasets, and a subset of the Sintel dataset with scenes containing humans. On all three datasets, we show that our method improves METRO and RAFT significantly, achieving state-of-the-art results in terms

of both human pose estimation and human optical flow estimation.

To summarize, the main contributions of our work are as follows:

- to the best of our knowledge, our work is the first to propose a multi-task-based inference-time optimization framework that enhances both human pose and optical flow estimation;
- we achieve the state-of-the-art human pose estimation performance on the Human 3.6M and the 3DPW datasets;
- we achieve state-of-the-art human optical flow performance for the Human 3.6M, the 3DPW, and the Sintel (human subset) datasets.

2 Related works

Monocular 3D human pose estimation. Work on monocular 3D human pose estimation can be primarily divided into two categories. The first class of methods detect 2D keypoints—for example, such as ones provided by Deep Dual Consecutive Network (DCPose) [20]—and then lift them to their 3D counterparts [22, 25, 43]. While they differ in detail, these lifting processes are typically performed through a Neural Network that is trained from data.

Although lifting-based pose estimators provide highly accurate pose estimates on standard benchmark datasets such as Human3.6M [12], in the case of datasets that are closer to 'in-the-wild' setups [41] they do not perform as well. Moreover, their performance is highly dependent on the quality of the initial detection of 2D keypoints, as if 2D keypoints are wrong, there is often no way of recovery.

Another class of methods utilize a parametric body model, typically the skinned multi-person linear model (SMPL) [21] and their variants [26, 28]. Parametric body model-based methods utilize this model in various ways. Bogo et al. [2] iteratively optimizes the body model parameters to fit the 2D observations. Kanazawa et al. [15] regresses the model parameters with a Neural Network, given the input image. VIBE [17] has recently shown that robust video-based pose estimation is possible by incorporating a motion prior learned in an adversarial setup. These methods, however, fail in cases of heavy occlusion, fast motion, and multi-person occlusion, as we will show in our experiments.

Recently for occlusion, Lin et al. [19] introduced a transformer-based pose estimation approach (METRO) that performs well on the Human3.6M and 3DPW datasets. The main benefit of this method is that, by utilizing transformers, the method learns the relationship among the vertices on the human mesh model, thus the method is able to figure out where the human body parts are, even in the presence of occlusions. Chen et al. [4] takes motivation from the anatomy of the human skeleton and break down the task into the bone direction and bone length predictions after which the 3D joint positions are estimated. Lastly, [5, 29, 44] explore optical flow to improve human pose, but not the other way around.

Tangent to the aforementioned research direction of having a better pose estimator, Arnab et al. [1] focuses on improving what is already available. They take advantage of the fact that, within neighbouring video frames, multi-view information exists in the presence of camera motions, and enhance pose estimation results via optimizing them according to predefined criteria—matching 2D and 3D pose estimates, enforcing temporal consistency, and utilizing human pose priors. As our method is also aiming to enhance pose and flow estimates via inference time optimization, these criteria can easily be included, which we do, except for human pose priors since it is often highly data dependent.

Optical flow methods. Recent methods use Deep Networks trained on large datasets for estimating optical flow. Flownet [7], Flownet 2.0 [11], and Spatial Pyramid Network [30]

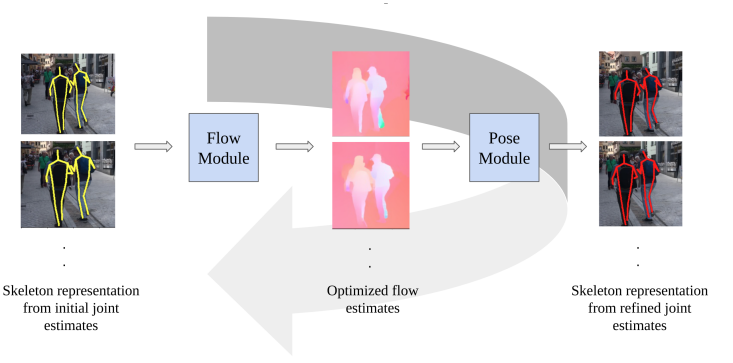


Figure 2: **Framework –** We use off-the-shelf human pose and optical flow estimators and further fine-tune their estimates to coincide with each other for improved performance.

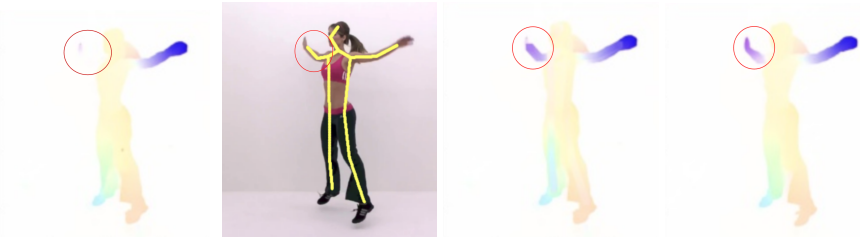
use two consecutive frames to estimate optical flow directly in an end-to-end manner, with varying architectural designs to make the algorithm more robust and efficient [11] or consider optical flow at various scale levels [30]. On the other hand, Voxel2Voxel [36] utilizes 3D convolutions to take both space and time into account through convolutions, so that more frames than just to two consecutive ones can be taken into account.

In addition to relying on the convolution structure directly, PWC-Net [34] utilizes a local cost volume approach, where all pixels in the potentially matching regions are compared with one another, hence explicitly forming relationships that are then utilized to create the optical flow map. This type of approach, where one allows the Deep Network to explicitly build relationship, has been highly effective in predicting optical flow [14, 37]. For example, state-of-the-art methods like Recurrent All-Pairs Field Transforms (RAFT) [33] and [13], propose to also utilize cost volumes, together with a recurrent setup that corrects optical flow estimate within the network. Recently, Correspondence Transformers (COTR) [14] further incorporates the cost volume and the recurrent strategy using Transformers [40].

In case of the performance of these methods on estimating human optical flow, however, is somewhat questionable [31, 32]. Optical flow methods are trained with scenes without particular focus on humans. These include real-world driving datasets such as KITTI [8], and synthetic datasets such as FlyingChairs [7], FlyingThings [23], and Sintel [3]. Because of this, many of them perform poorly when applied to human centred tasks [31, 32]. Thus, there is room for improvement here, which in our work, we bring by utilizing human pose estimators—human optical flow, should human pose estimators be perfect, can be obtained by simply rendering the poses in the two frames from which we wish to extract optical flow.

3 Method

The overall framework is shown in Figure 2. We utilize off-the-shelf pose and flow estimation modules, namely METRO [19] for the 3D human pose estimates, DCPose [20] for the location of 2D joints, and RAFT [33] for flow estimation. With these pre-trained modules, we first obtain the human pose estimates within the scene, with which we create a rough human skeleton-based sketch of the human optical flow. We then provide this sketch to RAFT, and fine-tune RAFT parameters for each scene so that the estimated flow follows this rough sketch flow. This already results in an improved optical flow estimate compared to simply



(a) Flow from [35] (b) Pose-based sketch (c) Target flow (d) Fine-tuned flow

Figure 3: **Example improving human optical flow –** With flow from off-the-shelf RAFT [35] we augment it with the pose estimate-based human optical flow to obtain a target optical flow which we fine-tune RAFT with for improved performance (highlighted circle).

using RAFT off-the-shelf. We use this optimized flow estimate and additional human pose related priors to then refine the pose estimates, again by fine-tuning the pose estimates directly on this scene. Finally, this process is repeated with the enhanced estimates. We detail each part of the pipeline in the following subsections.

3.1 Improving human optical flow with pose

To improve human optical flow estimates, we rely on human pose estimation. As shown in Figure 3 (a), as RAFT [35] is a generic optical flow estimator, the estimated flow may be inaccurate, especially when a human in the scene is moving abruptly. The human pose estimator, for example METRO [19], however, may still provide reasonable (albeit imperfect) pose estimates as shown in the skeleton sketch in Figure 3 (b). Hence, allowing the pose estimates to help the optical flow estimation process is a natural choice.

Specifically, for each frame, we apply the 3D human pose estimator, for example, METRO, to get the 3D joint locations. Next, we project the 3D joints onto the 2D image and create a stick-man figure as shown in Figure 3 (b), using the joint pairs as in [20] with the exception of the (head, neck) bone being replaced with the (head, nose) and (nose, neck) bones to roughly model also the facial motion. In addition, we make the skeleton thick so that it covers more than just a simple thin line by convolving a cross shape with a fifteen-pixel radius. When 3D joint estimates are unreliable, we directly use 2D joint estimates instead, for example from DCPose. We then compute the optical flow on the stick-man pixels trivially by taking the difference between stick-man pixel positions in consecutive frames. This creates a very rough optical flow map of the 'bones' of the human body. Next, the rough optical flow map of the bones is overlaid on top of the estimated flow map (*e.g.*, by RAFT as shown in Figure 3 (a)). This then leads to a 'target' flow map as shown in Figure 3 (c).

Finally, to leverage the implicit bias of optical flow estimates already stored in RAFT, and at the same time enhance the estimates, we fine-tune the RAFT network to generate this 'target' flow map. As fine-tuning for too long would lead to the RAFT network generating an output identical to the 'target' flow map, we fine-tune only for a very few iterations, which leads to an improved optical flow map as shown in Figure 3 (d). This somewhat resembles how Deep Image Prior [39] achieves image enhancement via early stopping an overfitting process. Here, similar to the Deep Image Prior, we are taking advantage of existing networks, and the learned prior about the task within them. Note that this early stopping also prevents our flow estimates from diverging too much due to potentially faulty estimates.

Mathematically, denoting the flow map prior to the optimization run as $\mathcal{F}^t$ where t is the frame index, the target flow generated as $\hat{\mathcal{F}}^t$, smooth ℓ_1 norm as ρ, and the parameters of the RAFT network as Φ_{RAFT} we minimize

$$\mathcal{L}_{\text{flow}}\left(\Phi_{\text{RAFT}}\right) = \mathbb{E}_t\left[\rho\left(\mathcal{F}^t - \hat{\mathcal{F}}^t\right)\right]. \tag{1}$$

3.2 Improving human pose with optical flow

With improved optical flow, we further improve our human pose estimates. Specifically, we enforce that the movement of the 3D joints, when projected onto the 2D image, follows the optical flow estimates. In addition to optical flow, we further enforce temporal consistency, as in [1], of the joint and camera estimates, including the human bone length inspired by [12], and also enforce that the results of the 2D joint estimator (DCPose) match that of the 3D method. Thus, if we denote the 3D joint locations as $\mathbf{X}$ and the corresponding camera estimates as $\mathbf{C}$, we minimize

$$\mathcal{L}_{\text{pose}}\left(\mathbf{X},\mathbf{C}\right) = \mathcal{L}_{\text{opt}}\left(\mathbf{X},\mathbf{C}\right) + \mathcal{L}_{\text{3D}}\left(\mathbf{X}\right) + \mathcal{L}_{\text{2D}}\left(\mathbf{X},\mathbf{C}\right) + \mathcal{L}_{\text{temp}}\left(\mathbf{X},\mathbf{C}\right). \tag{2}$$

Optical flow consistency—$\mathcal{L}_{\text{opt}}\left(\mathbf{X},\mathbf{C}\right)$. We make sure that the 2D projections of the 3D joints obey the optical flow at the joint locations as the two should be estimating the same phenomenon, just in two different ways. We thus penalize any difference between the two. Denoting the projection operation as $\mathcal{P}$, the optical flow at a pixel location $\mathbf{x}$ as $\mathcal{F}_{\mathbf{x}}$, and the j-th 3D joint location for the t-th frame as $\mathbf{X}_j^t$ we write

$$\mathcal{L}_{\text{opt}}\left(\mathbf{X},\mathbf{C}\right) = \lambda_{\text{opt}}\mathbb{E}_{t,j}\left[\rho\left(\mathcal{F}^{t-1}_{\mathcal{P}\left(\mathbf{X}_j^{t-1}\right)} - \left(\mathcal{P}\left(\mathbf{X}_j^t,\mathbf{C}^t\right) - \mathcal{P}\left(\mathbf{X}_j^{t-1},\mathbf{C}^{t-1}\right)\right)\right)\right]. \tag{3}$$

3D joint consistency—$\mathcal{L}_{\text{3D}}\left(\mathbf{X}\right)$. To prevent our optimized poses from deviating too much from the original estimate from the off-the-shelf method, we penalize when the deviation is too large. Denoting the initial estimate for $\mathbf{X}_j^t$ as $\tilde{\mathbf{X}}_j^t$ we write

$$\mathcal{L}_{\text{3D}}\left(\mathbf{X}\right) = \lambda_{\text{3D}}\mathbb{E}_{t,j}\left[\rho\left(\mathbf{X}_j^t - \tilde{\mathbf{X}}_j^t\right)\right]. \tag{4}$$

2D joint consistency—$\mathcal{L}_{\text{2D}}\left(\mathbf{X},\mathbf{C}\right)$. 2D joint estimates from DCPose are typically very accurate, and often more reliable than the 3D pose estimates, which is unsurprising given that estimating 3D pose introduces an additional dimension to the problem. Hence, we penalize when the 3D estimates deviates from them. Denoting the detection result from DCPose (or any other 2D human joint detector) as $\mathbf{x}$, and the projection by $\mathcal{P}$, we write

$$\mathcal{L}_{\text{2D}}\left(\mathbf{X},\mathbf{C}\right) = \lambda_{\text{2D}}\mathbb{E}_{t,j}\left[w_j^t\rho\left(\mathbf{x}_j^t - \mathcal{P}\left(\mathbf{X}_j^t,\mathbf{C}^t\right)\right)\right], \tag{5}$$

where w denotes the confidence score for a joint estimate coming from the joint detector.

Temporal consistency—$\mathcal{L}_{\text{temp}}\left(\mathbf{X},\mathbf{C}\right)$. As in [1], we leverage the fact that temporal consistency can be assumed, as change is small between frames. Unlike [1], we further enforce the bone length constraint when considering this temporal consistency term. We thus write

$$\begin{aligned}\mathcal{L}_{\text{temp}}\left(\mathbf{X},\mathbf{C}\right) = \lambda_{\text{pos}}\mathbb{E}_{t,j}\left[\rho\left(\mathbf{X}_j^t - \mathbf{X}_j^{t-1}\right)\right] + \lambda_{\text{cam}}\mathbb{E}_{t,j}\left[\rho\left(\mathbf{C}^t - \mathbf{C}^{t-1}\right)\right] \\ + \lambda_{\text{bone}}\mathbb{E}_{t,j,k}\left[\rho\left(\left\|\mathbf{X}_j^t - \mathbf{X}_k^t\right\|_2 - \left\|\mathbf{X}_j^{t-1} - \mathbf{X}_k^{t-1}\right\|_2\right)\right]\end{aligned}, \tag{6}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm.

3.3 Implementation details

Optimization settings. To optimize the flow network, we use the Adam optimizer [16] with a learning rate of 10^{-5} and default parameters. For the flow network (RAFT), we optimize its parameters for the entire video, per Eq. (1). This effectively results in an improved RAFT model for each scene. In more detail, we iterate over each frame one-by-one (equivalent to batch size of one) and optimize the per-scene RAFT network for eight epochs, a value that we empirically found that works well in general.

For optimizing the 3D joint estimates, we, again, use the Adam optimizer [16] with a learning rate of 0.001. We empirically set the number of optimization iterations to 1,500 epochs. We perform this flow and pose optimization cycle once for pose and twice for flow. See `Supplementary Material` for experiments regarding this choice.

Finally, for our experiments with METRO, we set $\lambda_{\text{opt}} = 0.01$, $\lambda_{\text{3D}} = 400$, $\lambda_{\text{2D}} = 0.01$, $\lambda_{\text{pos}} = 300$, $\lambda_{\text{cam}} = 0.1$, and $\lambda_{\text{bone}} = 10^4$, which we found empirically by testing a few videos that this setup works well in general.

When 3D joint estimates are unreliable. We found that METRO, our choice of the 3D joint estimator, does not perform as well when applied to datasets that are not focusing on humans. This leads into $\mathbf{X}$ being erroneous. In this case, we simply resort to the 2D joint detector DCPose, which delivers more robust performance in more complicated scenarios. We thus modify our losses, specifically directly replace the projected points $\mathcal{P}\left(\mathbf{X}_j^t, \mathbf{C}^t\right)$ with $\mathbf{x}^t$ for all losses. We also ignore loss components that directly use $\mathbf{X}$ and not $\mathcal{P}\left(\mathbf{X}_j^t, \mathbf{C}^t\right)$ and replace the 3D joint consistency with a 2D version using initial 2D estimates. Specifically, this setup is for the Sintel Human subset dataset, which we will discuss more in Section 4.1. With this setup, we optimize for 50 epochs per cycle for the flow instead of the eight previously. The setup for the joints remain the same, although we are now in 2D.

4 Results

For evaluation, we use the Human3.6M [12], 3DPW [41] and a subset of the Sintel (final) [3] datasets. In this section, we show that our method outperforms the state of the art human pose and optical flow results on all the three datasets, without introducing any retraining or additional datasets.

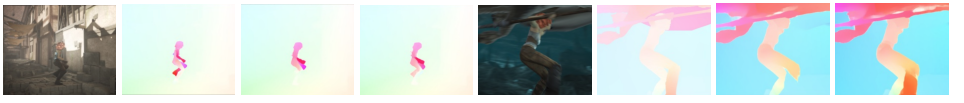
4.1 Datasets and evaluation setup

Datasets. We evaluate both the performance of human pose estimation and human optical flow estimation on three datasets—two aimed at human pose estimation and another at optical flow.

- **Human3.6M [12]:** This dataset is a large scale human pose dataset with 2D and 3D annotations captured in an indoor setting. There are a total of 2 subjects, S9 and S11, performing different actions such as walking and sitting. We follow the exact same evaluation setup as in METRO [19], based on their public implementation. As in [19] and [27], we down-sample the videos from 50 fps to 10 fps.

(a) METRO (b) Ours (c) METRO (d) Ours (e) DCPose (f) Ours (g) DCPose (h) Ours

Figure 4: **Pose estimation examples –** Example comparison of pose estimation results on the (a–b) Human3.6m, (c–d) on the 3DPW, and (e–h) on the Sintel human subset dataset. Note how our method improves pose estimates in case of occlusions (c–d) or when the person is interacting with the object (e–h).



(a) Input (b) GT (c) RAFT [43] (d) Ours (e) Input (f) GT (g) RAFT [43] (h) Ours

Figure 5: **Flow estimation examples –** Example comparison of optical flow estimation results on the Sintel human subset dataset. Ours successfully recovers the left leg in (d) and the right leg in (h), which was wrongly estimated in the case of the original RAFT method.

- **3DPW [41]:** This dataset is another large-scale dataset with 3D annotations of people *in the wild*. We use the standard test set of the 3DPW dataset—a total of 35K frames.
- **Sintel human subset:** We create a subset of the Sintel (final) dataset [3]—a commonly used optical flow dataset of synthetic scenes with atmospheric effects—by selecting scenes with humans in them: 'alley 2', 'bamboo 2', 'cave 2', and 'cave 4'. This results in a small dataset consisting of 196 frames with which we evaluate the human optical flow.

Metrics. We evaluate both the joint positions and the human optical flow errors with standard metrics. We use mean per joint position error (MPJPE) and the end point error (EPE). For the Human3.6M dataset and the 3DPW dataset we report the EPE values only for the joint locations where ground-truth correspondences are available—we do not know the ground-truth flow for all other points. For the Sintel dataset we use all points. In addition, we follow the standard method of computing MPJPE for the fourteen common joints.

4.2 Experimental results

Qualitative results. In Figure 4, we show qualitative results for how human pose estimation improves with our method. Note how the pose estimation result of existing methods can lead to partial inaccuracies, for example, due to (a–b) complex poses, (c–d) occlusions with other entities in the scene and (e–h) object interactions. Our method provides accurate results even in these cases. In Figure 5, we show results for improving flow estimates. As shown, optical flow values on the human that were originally missing and wrongly estimated are corrected with our method.

Quantitative results. In Table 1, we report the MPJPE and the EPE metrics for our method as well as other baselines. Building upon METRO and RAFT, our method outperforms both methods in their respective tasks. It is worth noting that, as shown earlier in Figure 4, the gains are most prominent when extreme poses occur and occlusions exist. Still, they are

Model	Human3.6M		3DPW		Sintel human	
	MPJPE↓	EPE↓	MPJPE ↓	EPE ↓	MPJPE ↓	EPE ↓
HMR [ ]	88.00	-	-	-	-	-
Pose2Mesh [ ]	64.90	-	89.20	-	-	-
I2LMeshNet [ ]	64.90	-	93.20	-	-	-
VIBE [ ]	65.60	-	82.00	-	-	-
PARE [ ]	-	-	74.50	-	-	-
METRO [ ]	54.04	-	77.10	-	-	-
METRO*	54.07	-	75.87	-	-	-
RAFT [ ]	-	2.729	-	1.751	-	2.513
Our method	**53.15**	**2.402**	**74.45**	**1.661**	-	**2.383**

Table 1: **Quantitative results –** Comparison with other 3D pose and optical flow methods multiple datasets. METRO* represents our adaptation of [ ] based on the official public implementation. Our method allows improving upon the state of the art, both in terms of human optical flow and pose.

Model	Human3.6M	
	MPJPE↓	EPE↓
METRO	54.04	-
METRO*	54.07	-
Anatomy3D	47.90	-
Avg(Anatomy3D, METRO)	43.34	-
RAFT	-	2.729
GMA	-	2.740
Avg(GMA, RAFT)	-	2.688
Our method – METRO / RAFT	53.15	2.402
Our method – METRO / GMA	53.14	2.316
Our method – Anatomy3D / RAFT	46.93	2.324
Our method – Anatomy3D / GMA	46.93	2.250
Our method – Avg(Anatomy3D, METRO)/Avg(GMA, RAFT)	**42.56**	**2.230**

Table 2: **Refinement on top of different 3D pose and optical flow methods** – A performance comparison on the Human3.6M dataset showing that our method is applicable to various methods.

frequent and large enough to be seen in the average metrics that we report in this table.

To further show that our optimization framework is not limited to METRO and RAFT, we conduct additional experiments with other optical flow and human pose estimators: GMA [ ], Anatomy3D [ ]. We also use them together with RAFT and METRO by averaging their estimates, which we found to work well.

We note that Anatomy3D does not provide camera estimates, and is designed to utilize the ground-truth camera parameters. As such, when Anatomy3D is used, we rely on ground-truth cameras. In addition, Anatomy3D estimates are initially more accurate than those from METRO, rendering the temporal consistency term useless. We thus do not use this term when Anatomy3D estimates are used.

We report these results in the Table 2—our method enhances *all* methods, demonstrating its efficacy.

Ablation study. We provide an ablation study on the losses and the number of optimization cycles in the `Supplementary Material`.

5 Conclusions

We proposed an iterative framework for enhancing human pose and optical flow estimation accuracy of existing methods *without* any training. Our method takes its roots from the fact that, when performed properly, the two tasks should coincide. Hence, we optimize optical flow to match the movement of the human joints and vice versa. This leads to a performance boost, enough to push the boundaries of the state of the art further. The gain was especially visible in cases where extreme poses, occlusions, and object-human interactions exist. We have validated our method on two human pose datasets, Human3.6M and 3DPW, and a subset of the Sintel optical flow dataset, achieving state of the art in all three datasets, both in terms of human pose and optical flow estimation accuracy.

Limitations and future work. The performance of our framework, while it improves upon the state of the art, is also somewhat bound by the quality of the initial estimates. Hence, starting completely from a wrong pose would not lead to accurate pose estimates, even with good optical flow. The opposite is also true. Thus, a promising research direction would be

to include multiple methods into the framework, thus reducing the risk of total failure. A naive version of this was shown in Table 2 via averaging, but a more sophisticated way of combining methods, for example via a multiple-instance learning setup could be interesting.

Additionally, the inference time of our pipeline is a limitation as we are, in fact, optimizing a network during inference time—it takes 20.50 seconds per frame on a GeForce RTX 3090 GPU using METRO and RAFT. This is roughly 14.5 times slower than running METRO and RAFT separately. Speeding up this inference time could be an interesting future research direction.

References

[1] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019.

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016.

[3] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pages 611–625. Springer, 2012.

[4] Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-Aware 3D Human Pose Estimation with Bone-Based Pose Decomposition. *IEEE Trans. Circuits Syst. Video Technol.*, 32(1):198–209, 2021.

[5] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. Occlusion-aware networks for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 723–732, 2019.

[6] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision*, pages 769–787. Springer, 2020.

[7] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015.

[8] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11): 1231–1237, 2013.

[9] Joseph Gesnouin, Steve Pechberti, Guillaume Bresson, Bogdan Stanciulescu, and Fabien Moutarde. Predicting intentions of pedestrians from 2d skeletal pose sequences with a representation-focused multi-branch deep learning network. *Algorithms*, 13(12): 331, 2020.

[10] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–84, 2018.

[11] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[12] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 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, 2013.

[13] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to Estimate Hidden Motions with Global Motion Aggregation. In *CVPR*, 2021.

[14] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. COTR: Correspondence Transformer for Matching Across Images. In *ICCV*, 2021.

[15] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018.

[16] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2014.

[17] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *CVPR*, pages 5253–5263, 2020.

[18] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021.

[19] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, 2021.

[20] Zhenguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, and Xun Wang. Deep dual consecutive network for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 525–534, 2021.

[21] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[22] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, pages 2640–2649, 2017.

[23] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016.

[24] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020.

[25] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2823–2832, 2017.

[26] Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)*, pages 598–613, 2020. URL https://star.is.tue.mpg.de.

[27] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017.

[28] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[29] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 1913–1921, 2015.

[30] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017.

[31] Anurag Ranjan, Javier Romero, and Michael J Black. Learning human optical flow. *arXiv preprint arXiv:1806.05666*, 2018.

[32] Anurag Ranjan, David T Hoffmann, Dimitrios Tzionas, Siyu Tang, Javier Romero, and Michael J Black. Learning multi-human optical flow. *International Journal of Computer Vision*, 128(4):873–890, 2020.

[33] Jan Stenum, Kendra M Cherry-Allen, Connor O Pyles, Rachel D Reetzke, Michael F Vignos, and Ryan T Roemmich. Applications of pose estimation in human health and performance across the lifespan. *Sensors*, 21(21):7315, 2021.

[34] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018.

[35] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

[36] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Deep end2end voxel2voxel prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 17–24, 2016.

[37] Prune Truong, Martin Danelljan, and Radu Timofte. GLU-Net: Global-local universal network for dense flow and correspondences. In *CVPR*, 2020.

[38] Amin Ullah, Khan Muhammad, Javier Del Ser, Sung Wook Baik, and Victor Hugo C de Albuquerque. Activity recognition using temporal optical flow convolutional features and multilayer lstm. *IEEE Transactions on Industrial Electronics*, 66(12):9692–9702, 2018.

[39] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *NIPS*, 2017.

[41] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018.

[42] Zhenbo Yu, Bingbing Ni, Jingwei Xu, Junjie Wang, Chenglong Zhao, and Wenjun Zhang. Towards alleviating the modeling ambiguity of unsupervised monocular 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8651–8660, 2021.

[43] Ruiqi Zhao, Yan Wang, and Aleix M Martinez. A simple, fast and highly-accurate algorithm to recover 3d shape from 2d landmarks on a single image. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):3059–3066, 2017.

[44] Shihao Zou, Chuan Guo, Xinxin Zuo, Sen Wang, Pengyu Wang, Xiaoqin Hu, Shoushun Chen, Minglun Gong, and Li Cheng. Eventhpe: Event-based 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10996–11005, 2021.