Bootstrapping Human Optical Flow and Pose

Aritro Roy Arko, James J. Little, Kwang Moo Yi
University of British Columbia

Problem:

→ Generic optical flow methods (such as RAFT) perform better on humans when fine-tuned on human-centric scenes. In addition, they fail in cases of fast motion.

RAFT

→ Overlooked assumption in recent pose estimations works (such as METRO) is temporal consistency. Some methods take them into consideration but most leave it for the Neural Network to implicitly figure out and embed into the framework while training.

METRO

Solution:

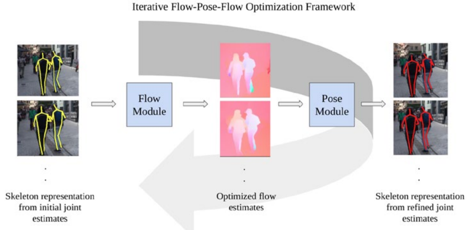
→ Make use of the tools that already exist—human pose estimators and optical flow networks—and enhance their performance by marrying the two.

→ Create an iterative flow-pose-flow optimization framework for inference.

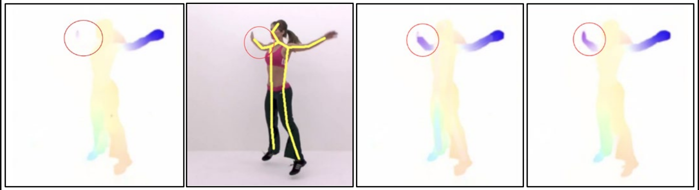
→ Idea originates from the fact that the movement of the joints, when projected in 2D, should follow the optical flow estimates at these locations.

METRO estimates | RAFT estimates | Flow enhanced pose | Pose enhanced flow

Overall framework:



Flow module:



Flow from RAFT | Pose-based sketch | Target flow | Fine-tuned flow

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

→ Generate rough optical flow map of the bones with help of pose estimator.

→ This is overlaid on top of the estimated flow map (e.g., by RAFT).

→ Target flow map ($\hat{\mathcal{F}}^t$) produced.

→ Minimize smooth ℓ_1 norm (ρ) between predicted flow ($\mathcal{F}^t$) and $\hat{\mathcal{F}}^t$. Update parameters of the RAFT model (Φ_{RAFT}) to get fine-tuned optical flow.

Pose module:

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

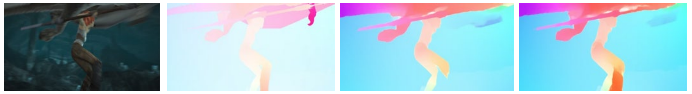
We directly optimize the 3D joint estimates based on optical flow consistency ($\mathcal{L}_{\text{opt}}(\mathbf{X}, \mathbf{C})$), 3D joint consistency ($\mathcal{L}_{\text{3D}}(\mathbf{X})$), 2D joint consistency ($\mathcal{L}_{\text{2D}}(\mathbf{X}, \mathbf{C})$) and temporal consistency ($\mathcal{L}_{\text{temp}}(\mathbf{X}, \mathbf{C})$).

Results:

Model	Human3.6M MPJPE↓	Human3.6M EPE↓	3DPW MPJPE↓	3DPW EPE↓	Sintel human MPJPE↓	Sintel human EPE↓
HMR [15]	88.00	-	-	-	-	-
Pose2Mesh [6]	64.90	-	89.20	-	-	-
I2LMeshNet [24]	64.90	-	93.20	-	-	-
VIBE [17]	65.60	-	82.00	-	-	-
PARE [18]	-	-	74.50	-	-	-
METRO [19]	54.04	-	77.10	-	-	-
METRO*	54.07	-	75.87	-	-	-
RAFT [35]	-	2.729	-	1.751	-	2.513
Our method	**53.15**	**2.402**	**74.45**	**1.661**	-	**2.383**

Model	Human3.6M MPJPE↓	Human3.6M 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**

DCPose | Ours | DCPose | Ours | METRO | Ours



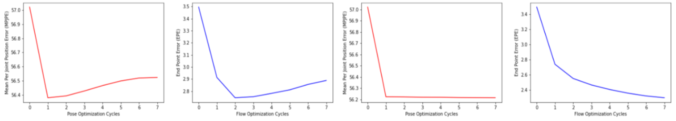
Input | Ground Truth | RAFT | Ours

Input | Ground Truth | RAFT | Ours

Ablation Study:

Method	MPJPE ↓
Initial pose estimates (METRO)	54.07
$\mathcal{L}_{\text{3D}}$	54.07
$\mathcal{L}_{\text{3D}} + \mathcal{L}_{\text{2D}}$	53.93
$\mathcal{L}_{\text{3D}} + \mathcal{L}_{\text{2D}} + \mathcal{L}_{\text{temp}}$ (without bone consistency)	53.45
$\mathcal{L}_{\text{3D}} + \mathcal{L}_{\text{2D}} + \mathcal{L}_{\text{temp}}$	53.29
$\mathcal{L}_{\text{3D}} + \mathcal{L}_{\text{2D}} + \mathcal{L}_{\text{temp}} + \mathcal{L}_{\text{opt}}$	**53.15**

→ Effects of adding different loss terms to our pose refinement pipeline.



(a) MPJPE vs #cycles (b) EPE vs #cycles (c) MPJPE vs #cycles /w GT (d) EPE vs #cycles /w GT

→ Pose and flow errors with respect to the number of optimization cycles.