Enhancing Human Optical Flow via 3D Spectral Prior

Shiwei Mao, Mingze Sun and Ruqi Huang†

Tsinghua Shenzhen International Graduate School, China

Abstract

In this paper, we consider the problem of human optical flow estimation, which is critical in a series of human-centric computer vision tasks. Recent deep learning-based optical flow models have achieved considerable accuracy and generalization by incorporating various kinds of priors. However, the majority either rely on large-scale 2D annotations or rigid priors, overlooking the 3D non-rigid *nature of human articulations. To this end, we advocate enhancing human optical flow estimation via* 3D spectral prior-aware *pretraining, which is based on the well-known functional maps formulation in 3D shape matching. Our pretraining can be performed with synthetic human shapes. More specifically, we first render shapes to images and then leverage the natural inclusion maps from images to shapes to lift 2D optical flow into 3D correspondences, which are further encoded as functional maps. Such lifting operation allows to inject the intrinsic geometric features encoded in the spectral representations into optical flow learning, leading to improvement of the latter, especially in the presence of non-rigid deformations. In practice, we establish a pretraining pipeline tailored for triangular meshes, which is general regarding target optical flow network. It is worth noting that it does not introduce any additional learning parameters but only require some pre-computed eigen decomposition on the meshes. For RAFT and GMA, our pretraining task achieves improvements of* 12.8% *and* 4.9% *in AEPE on the SHOF benchmark, respectively.*

CCS Concepts

• ***Computing methodologies*** → ***Spectral methods;***

1. Introduction

Optical flow, defined as a vector field that represents the velocity or displacement of each pixel in an image, is a fundamental and crucial concept in computer vision. It is widely applied in various domains such as motion analysis, video stabilization [LTC*21], object tracking [KPD15], and 3D reconstruction [LZYX22]. In this paper, we particularly focus on estimating optical flows on images involving humans, which is critical in a series of human-centric computer vision tasks, including human mesh recovery [LL21], pose estimation [LXH*22, JYSP21], and 3D human tracking [SBL*23].

Classical optical flow estimation is cast as an optimization problem based on brightness constancy and spatial smoothness assumptions. However, they may fail due to illumination changes or complex lighting conditions. To tackle such kind of problem, a series of deep learning methods [IMS*17, SYLK18, TD20, LYL*22] have been proposed to exploit the strong feature-extracting capacity of neural networks as well as accumulated data annotations. Apart from the above supervised methods, a notable recent trend is to pretrain feature extractors on large-scale pretraining models to further enhance learning-based optical flow estimation, such as image matching [DCF23], image classification [HSZ*22, SHL*23, SHB*23], and masked image modeling [WLL*22, WLL*23a]. Another line of work makes use of 3d information by jointly optimizing the 2D optical flow and 3D scene flow [LLX*22, TD21, LZL*22]. However, all the above work either relies on large-scale 2D annotations or rigid priors, which falls short of accurately characterizing the *complex, non-rigid* 3D human articulation. In particular, we observe that current methods may make mistakes on frames with a large temporal gap since it is nontrivial to find the precise corresponding pixels under large deformation.

In light of this, we draw inspiration from the well-established functional maps formulation [OBCS*12]. It is originally proposed for 3D non-rigid shape matching, where shapes may undergo significant articulations. More specifically, we compute eigenbasis and eigenvalues of 3D shapes to encode the respective intrinsic geometry, and treat functional maps as transformation matrices between eigenbases of shapes induced by the point-wise correspondences. The algebraic structure of functional maps is elegantly related to the underlying maps' properties, such as isometry, area-preserving, and conformal [OBCS*12].

Our main goal then turns into associating 2D optical flow and 3D shape correspondences. Noting that images are essentially projections of 3D shapes from certain viewpoint, the association can

† Corresponding author: ruqihuang@sz.tsinghua.edu.cn

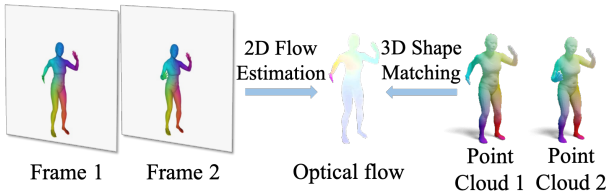

Figure 1: *For each frame, during the mesh rendering, we can calculate the corresponding point cloud coordination for each pixel. The shape matching correspondence can be transformed into optical flow and vice versa. Thus, we can use the spectral prior of the point clouds to constrain the optical flow estimation between the two frames.*

be readily done via the natural inclusion maps from rendered images to the regarding 3D shapes. In particular, given a set of synthetic 3D shapes represented as triangle meshes, we first render them into images, and then propose a formulation that transfers optical flow into correspondences between shapes. Now we can then leverage the rich 3D information encoded in eigenbasis to further enhance the optical flow estimation. Due to the inherent connection between optical flow and 3D shape matching, the two formats can be mutually converted, which enables 3D shape matching to be a suitable pretraining task for enhancing non-rigid optical flow estimation. Moreover, recent advancement on Deep Functional Maps (DFM) [RSO19] allows for unsupervised shape matching by leveraging structural properties of functional maps, say, bijectivity or area-preserving, which are formulated as algebraic regularization terms. We also incorporate such regularization terms in our pretraining task, reducing the burden of data annotation. To conclude, we leverage the structural properties of functional maps to guide the image feature extractor.

In practical implementation, we realize a unique rendering module, which is specially designed for non-rigid human correspondence estimation. For each image, the corresponding point cloud is inherently partial. Thus, we propose a regularized functional maps formulation and all we need is the spectral features of the full mesh. Note that during the pretraining task, we do not introduce any learnable parameters and utilize the properties of functional maps to train the network in an unsupervised way. Since we translate the final per-pixel optical flow into functional maps, our pretraining task can be applied to most optical flow networks and experiments demonstrate that our pretraining task can indeed improve the accuracy and generalization of human optical flow estimation.

To summarize, our main contributions are as follows: (1) We combine optical flow estimation with non-rigid shape matching and leverage the 3D spectral information to enhance 2D information; (2) We design a novel pretraining task for human optical flow estimation and experiments demonstrate that our pretraining task can enhance the optical flow feature extractor; (3) We propose a mesh rendering component, which can be applied to any synthetic 3D shapes.

2. Related Works

Optical Flow: The global formulation of optical flow is first introduced by [HS81], which relies on both brightness constancy and spatial smoothness assumptions. Based on this, many improved traditional optical flow estimation methods [BA96, BBPW04] have been explored. After that, FlowNet [DFI*15] first utilizes a convolutional network and proposes a synthetic dataset, pioneering the optical flow estimation based on deep learning, such as [IMS*17, SYLK18, RB17, HR19, ZSD*20]. Subsequently, RAFT [TD20] uses recurrent update operators to iteratively refine the estimated flow, indicating strong generalization and accuracy. To tackle the occlusion problem, based on RAFT, GMA [JCL*21] proposes a global motion aggregation module to model image self-similarities. Inspired by these works, super kernels [SCZ*22], graph reasoning [LYL*22], equilibrium formulation [BGSK22], Gaussian attention [LYL*23], and backward accumulation [WLL*23b] are applied to further enhance the performance. Despite the convolutional network, [XZC*22, HSZ*22, SHL*23, SHB*23] employ transformers as the feature enhancement or feature extractor. Apart from supervised methods, [YHD16, MHR18, JSB*20] also seek for unsupervised methods based on photometric consistency or bidirectional flow estimation. Recently, various pretraining tasks have been developed to benefit the optical flow estimation, such as geometric image matching [DCF23], image classification [HSZ*22, SHL*23, SHB*23], and masked image modeling [WLL*22, WLL*23a].

Scene Flow: 3D scene flow can be estimated directly from two point clouds and the main purpose is to move the source point cloud to the target point cloud. RAFT-3D [TD21] proposes a single RAFT-based network with rigid motion embedding for scene flow estimation. [GLW*21] suggests predicting object-based flow instead of per-pixel flow by segmenting the scene into rigid and dynamic parts to ensure estimating consistent flow for each object. RigidFlow [LZL*22] uses over-segmentation to divide point clouds into different parts and generate rigid transformation for each supervoxel. In this way, RigidFlow generates pseudo labels for scene flow and realizes a self-supervised learning scheme. Although combining optical flow estimation and scene flow estimation can improve both two tasks, the network requires a feature extractor especially for point clouds or depth images, such as CamLiFlow [LLX*22]. Different from these methods, we utilize the spectral information of 3D representations and do not need to add other learning networks.

Human Optical Flow: The optical flow of humans is known to be useful, however, most optical flow dataset focus on rigid things, such as FlyingChairs [DFI*15] and FlyingThings3D [MIH*16]. Though Sintel [BWSB12] contains non-rigid motions, it has a limited number of synthetic scenes. In order to realize an optical flow estimator, especially for human motion, SHOF [RRB18] creates an extensive dataset containing images of realistic single human shapes in motion and demonstrates that optical flow methods trained on this dataset can generalize well to the real world human scenes. After that, MHOF [RHT*20] further designs a dataset for multi-human optical flow estimation. DeepDeform [BZTN20] proposes a semi-supervised strategy combining self-supervision with sparse annotations to build a large-scale RGB-D dataset of non-rigidly deforming scenes. The dataset provides optical flow ground truth for part of the frames.

Deep Functional Maps: A noticeable trend among the learning-

based shape matching approaches is based on the formalism of Deep Functional Maps (DFM), pioneered by the FMNet [LRR*17], Functional Maps [OBCS*12], as a spectral map representation, allows to encode maps into compact matrices and to express desirable map priors (e.g., area-preservation, isometry, bijectivity) in simple algebraic forms. Instead of learning from labeled maps, unsupervised approaches [HLR*19, RSO19] demonstrate that it is sufficient to learn from geometric map priors. More recent advances take advantage of the multi-scale properties of the eigenbasis of the Laplace-Beltrami operator. In the works [MRR*19, HRWO20, ELC20, ETLTC20], conversions are done between spatial domain and a series of spectral domains spanned by eigenfunctions of increasing dimensions. [SMJ*23] leverages the cycle consistency between spatial and spectral domains, further increasing the accuracy, consistency, and generalization performance. Recently, DFM has been applied to non-rigid shape registration [JSH23], which also inspires our work to combine DFM with optical flow estimation.

3. Methodology

It is known that 3D structures contain much richer geometry information than 2D images and can indeed assist or complement 2D information. For optical flow estimation, the common inputs are RGB images and we hope to utilize the ample 3D geometry feature to guide the 2D image feature. We find that DFM is widely used in non-rigid shape matching and the correspondence between 3D meshes is identical to the optical flow between their rendered images. Thus, we use non-rigid shape matching as the pretraining task, in order to leverage the useful 3D information to enhance the image feature extractor, especially for optical flow estimation.

3.1. Dataset Preparation

In the pretraining task, we use DFM to constrain the correspondence between different shapes thus enhancing the optical flow estimation. In order to realize this, we create a unique pretraining dataset, especially for non-rigid human correspondence estimation. We use D-FAUST [BRPMB17], the first 4D dataset providing both real scans and dense ground-truth correspondences between them, as the prior of the pretraining dataset. D-FAUST includes dynamic performances of 10 subjects of various shapes and ages and provides raw and aligned meshes. We randomly choose 20 sequences for training and 3 sequences for validation. For the purpose of predicting motion transformation precisely, we remove the ones with relatively slight movement. At last, we use 5,325 shapes as the training part and 673 as the validation one.

Mesh Preprocessing: For each mesh, we first normalize the surface area to one and then compute the leading k eigenfunctions of the Laplace-Beltrami operator on each shape, which can be treated as a high-dimensional spectral embedding of the respective shape, including eigenvectors and eigenvalues. Since the raw meshes lack color or texture, for the sake of helping the optical flow network to better utilize the color information, we add a color map as the texture of the raw meshes. To ensure the consistency between meshes in one sequence, we further add texture through the correspondences between the first shape and the other ones.

Mesh Rendering: The normalized meshes of D-FAUST contain

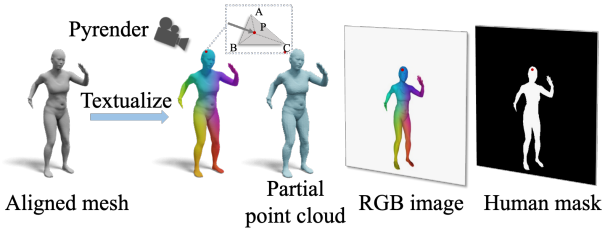

Figure 2: *The whole rendering process of the pretraining dataset. For each pixel in the rendered image, Pyrender creates a ray to hit the mesh and we calculate the point coordinates to combine the 2D pixels with 3D points.*

6,890 vertices and 13,776 faces. We use Pyrender to render the meshes and for one shape, we need a rendered RGB image, human mask information, a partial point cloud that corresponds to the masked pixels of the rendered image, and partial eigenvectors for the point cloud. During the rendering, as is shown in Fig. 2, we set the camera pose to a frontal view, and the camera perspective will be transformed into rays based on the predefined image resolution, which is 384×384. Each ray will record the face index it hits and calculate its corresponding point cloud coordinates on the original mesh. Through this way, we can obtain a point cloud that is identical to pixels in the rendered image and thus successfully combine the 2D image pixels with the 3D point cloud coordinates. As for handling human mask information, if the ray does not contain a face index, its value is 0, and otherwise, its value is 1. For each shape, the sum of the value 1 in the mask is the same as the number of the point cloud. According to the coordinates of the partial point cloud, we can calculate the eigenvectors by barycentric coordinates. The point is generated from a face which is a triangle mesh containing three vertices A, B, and C. Then, with barycentric coordinates, any point P can be uniquely expressed as:

$$P = uA + vB + wC, u + v + w = 1. \tag{1}$$

We refer readers to the Supplementary Material for more details on the barycenter computation.

Recall that the eigenvectors $\Phi \in \mathbb{R}^{n \times k}$. For the three vertices A, B, C, we let Φ_A, Φ_B, and Φ_C the regarding rows of *Phi*. Once we get the barycentric coordinates (u, v, w), we can calculate the eigen embedding for point P, Φ_P, as:

$$\Phi_P = u\Phi_A + v\Phi_B + w\Phi_C. \tag{2}$$

Similarly, we can obtain the eigen embeddings for all point P, which consists of $\Phi_{\text{pcd}} \in \mathbb{R}^{m \times k}$, where m is the number of points lifted from visible pixels. We can now combine the 3D spectral information with 2D images, which enables us to leverage 3D geometry to enhance optical flow estimation.

3.2. Main Pipeline

As shown in Fig. 1, our key insight is to lift 2D optical flow into 3D shape correspondences, which, when expressed as functional maps, can be regularized by mild and general geometric map priors. Sun *et al.* [SMJ*23] presents a novel design of unsupervised DFM, which effectively enforces the harmony of learned maps under the

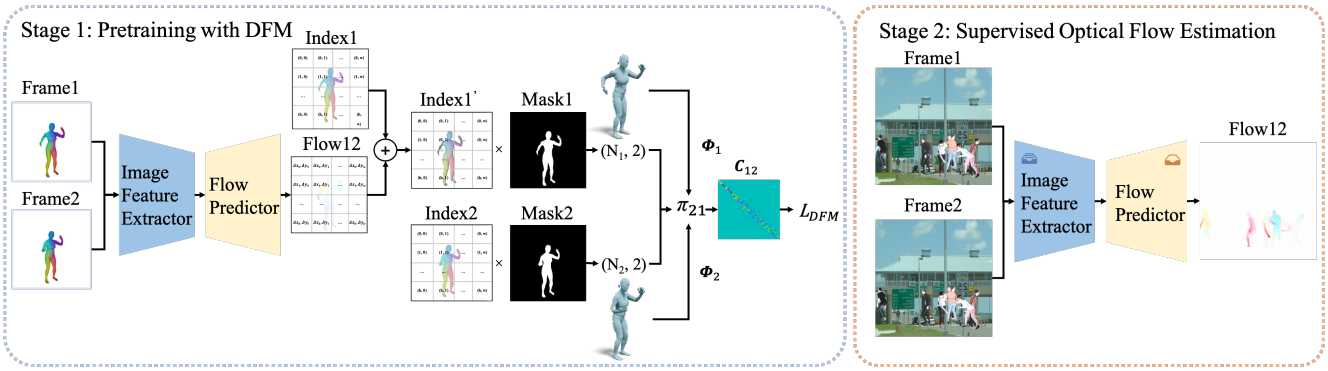

Figure 3: *The main framework can be divided into two stages. In the first stage, we transform the estimated optical flow into functional maps according to the coordinates and corresponding partial eigenvalues. We then use the functional maps penalties to constrain the optical flow and thus enhance the image feature extractor. In the second stage, we load the parameters of the image feature extractor and train the flow predictor from raw with supervised optical flow estimation.*

spectral and the point-wise representation. Inspired by this, we can view optical flow as the point-wise representation, also known as the spatial correspondence, and leverage the spectral consistency constrained by deep functional maps. In this way, we can make use of the 3D spectral features from meshes and enhance 2D optical flow estimation. Most optical flow networks can be divided into image feature extractors and optical flow estimators. Our main purpose is to pretrain a robust image feature extractor which is enhanced by 3D information and our pretraining task can be applied on a wide range of optical flow networks. Our main pipeline consists of the following three main components:

1. Lift Optical Flow to Functional maps: Let $\mathcal{I} = \{\mathbf{I}_1, \ldots, \mathbf{I}_N\}$ denote a sequence with N image frames $\mathbf{I}_t \in \mathbb{R}^{w\times h\times 3}$ of size $w \times h$ and 3 color channels and the indexes of each frame is $\mathbf{Index}_t \in \mathbb{R}^{w\times h\times 2}$. In the pretraining stage, we also have human mask information $\mathbf{M}_t \in \mathbb{R}^{w\times h\times 1}$ and eigenvectors $\Phi_t \in \mathbb{R}^{n_t\times k}$. Let $\mathbf{F}_{ij} \in \mathbb{R}^{w\times h\times 2}$ denote the optical flow field from the reference image $\mathbf{I}_i$ to the target image $\mathbf{I}_j$. Specifically, for each pixel $\mathbf{x} \in \Omega_i = \{1,\ldots,w\} \times \{1,\ldots,h\}$ in reference image $\mathbf{I}_i$, $\mathbf{F}_{ij}(\mathbf{x}) \in \mathbb{R}^2$ describes the apparent motion from frame $\mathbf{I}_i$ to $\mathbf{I}_j$.

The main framework can be divided into two stages, the pretaining stage and the finetuning stage. In stage one, we transform the estimated optical flow into point-wise correspondence which is then guided by deep functional maps. Given a source frame $\mathbf{I}_i$ and a target frame $\mathbf{I}_j$, through the feature extractor and flow estimator, we can get the optical flow $\mathbf{F}_{i,j}$, which means the offset between $\mathbf{I}_i$ and $\mathbf{I}_j$. In order to calculate the point-wise mapping, we first warp the indexes of $\mathbf{I}_i$ into $\mathbf{I}_j$ by adding $\mathbf{F}_{ij}$, which is formulated as:

$$\mathbf{Index}_i^{'} = \mathbf{Index}_i + \mathbf{F}_{ij}. \tag{3}$$

Next, we apply $\mathbf{M}_i$ on $\mathbf{Index}_i^{'}$ and $\mathbf{M}_j$ on $\mathbf{Index}_j$. Now, we get $\mathbf{P}_i \in \mathbb{R}^{n_i\times 2}$ and $\mathbf{P}_j \in \mathbb{R}^{n_j\times 2}$ of pixel numbers and 2 channels indicating the coordinates of the pixels. We can compute the soft point-wise map by nearest neighbor searching between the rows of $\mathbf{P}_i$ and those of $\mathbf{P}_j$. Given a pair of indices $p \in [1..n_i], q \in [1..n_j]$, we compute residual:

$$\delta_{qp} = \|\mathbf{P}_i[p] - \mathbf{P}_j[q]\|_2, \tag{4}$$

where $\mathbf{P}_i[p]$ denotes the p^{th} row of $\mathbf{P}_i$, and similarly we define $\mathbf{P}_j[q]$. The soft point-wise map $\Pi \in \mathbb{R}^{n_j\times n_i}$ is then given by:

$$\Pi_{ji}(q,p) = \frac{\exp(-\alpha\delta_{qp})}{\sum_{p'}\exp(-\alpha\delta_{qp'})}. \tag{5}$$

Note that by construction, each row of Π_{ji} is non-negative and sums up to 1, forming a probability distribution. The parameter α controls the entropy of each distribution – the smaller/larger α is, the fuzzier/sharper the distribution is. Instead of manually tuning the optimal α, following [SMJ*23], we use a learning scheme that dynamically controls α over training.

The next step is to convert the optical flow induced point-wise soft maps into functional maps representation [OBCS*12], so that we can leverage the rich structural properties of the latter. When dealing with maps between full shapes, the map conversion is straightforward. We refer readers to the Supplementary Material for a brief overview of functional maps formulation. However, the point clouds corresponding to the rendered images are inherently partial, necessitating a more specific treatment. In particular, we compute the functional map $\mathbf{C}_{ij}$ by optimizing the following energy:

$$\min_{\mathbf{C}_{ij}\in\mathbb{R}^{k\times k}} \|\Phi_j\mathbf{C}_{ij} - \Pi_{ji}\Phi_i\|^2 + \lambda\left\|\mathbf{C}_{ij}\Delta_i - \Delta_j\mathbf{C}_{ij}\right\|^2, \tag{6}$$

where Φ_i, Φ_j are the first k eigenvectors regarding all points lifted from rendering (Eqn. 2), and Δ_i, Δ_j are diagonal matrices of the first k eigenvalues on the input full shapes, and λ is a scalar regularization parameter.

2. Unsupervised Pretrain with Deep Functional Maps: The key idea of functional map representation is to encode shape correspondences as transformations between the respective spectral embeddings, which are represented by compact matrices by using reduced eigenbasis. [RSO19] introduces a novel approach for spectral unsupervised functional maps network, leveraging structural properties of the inferred functional maps, such as their bijectivity or orthogonality. These penalties can be applied during optimization purely in the spectral domain. As is illustrated above, the optical flow $\mathbf{F}_{ij}$ can be transformed into functional maps $\mathbf{C}_{ij}$ and $\mathbf{C}_{ji}$. We can opti-

mize the optical flow using an unsupervised DFM loss as the loss function and also propose an unsupervised scheme for optimizing optical flow. Below, we briefly introduce the unsupervised DFM loss.

Bijectivity: Given the functional maps in both directions, the simplest requirement is for them to be inverses of each other, which can be enforced by penalizing the difference between their composition and the identity map. The penalty used in [ERGB16] can be written as:

$$E_{\text{bij}} = \left\|\mathbf{C}_{ij}\mathbf{C}_{ji} - \mathbf{I}\right\|^2 + \left\|\mathbf{C}_{ji}\mathbf{C}_{ij} - \mathbf{I}\right\|^2. \tag{7}$$

Orthogonality: As observed in several works [OBCS*12, ROA*13], a point-to-point map is local area preserving if and only if the corresponding functional map is orthonormal. Thus, approximately satisfying this assumption, a natural penalty can be formulated as:

$$E_{\text{ortho}} = \left\|\mathbf{C}_{ij}^{\top}\mathbf{C}_{ij} - \mathbf{I}\right\|^2 + \left\|\mathbf{C}_{ji}^{\top}\mathbf{C}_{ji} - \mathbf{I}\right\|^2. \tag{8}$$

Based on the above two penalties, we can formulate the DFM loss as:

$$L_{\text{DFM}} = \lambda_{\text{bij}}E_{\text{bij}} + \lambda_{\text{ortho}}E_{\text{ortho}}, \tag{9}$$

where $\lambda_{\text{bij}} = \lambda_{\text{ortho}} = 1$. We use L_{DFM} as the loss function during our pre-training phase.

3. Finetuning: Built on the above pretraining process, we only load the parameters of the image feature extractor and train the flow predictor from raw (see the right panel of Fig. 3).

4. Experiments

4.1. Experimental Setting

Dataset: Since our pretraing method mainly focuses on non-rigid shape matching, we choose several non-rigid optical flow dataset to evaluate the performance. SHOF refers to the single human optical flow dataset, an extensive dataset containing images of realistic human shapes in motion together with ground truth optical flow. SHOF uses the SMPL model [LMR*23] to generate a wide variety of different human shapes and appearances. During the rendering process, real images are added as the background. In addition to the movement of human poses, SHOF also includes the overall displacement of the background image. The image resolution in SHOF is 256×256. We use 135,153 pairs of images as the training set, 530 pairs as validation, and 53,919 as the testing set. MHOF refers to the multi-human optical flow dataset which contains multiple people involving significant occlusion between them. In order to render more realistic images, MHOF replaces the SMPL model with SMPL+H [RTB22]. Virtual backgrounds are also added to support deep learning systems while the motions of the backgrounds are slighter than SHOF. Instead, MHOF pays more attention on human motions, which also makes it harder to make precise optical flow estimation. The image resolution in MHOF is 640×640. We use 86,218 pairs for training and 13,236 pairs for testing. DeepDeform [BZTN20] contains a large-scale RGB-D dataset of non-rigid scenes, including 400 scenes, 390,000 frames, and 5,533 densely aligned frame pairs. The image resolution is 480×640. DeepDeform contains optical flow ground truth for frames with large gaps.

Metrics: To assess the performance, we employ two metrics: the Average End-Point Error (AEPE) and the 1px, 3px, and 5px accuracy. AEPE measures the average error in the flow across all valid pixels while the pixel accuracy indicates the precision or accuracy with respect to 1px, 3px, and 5px. This dual metric approach allows for a comprehensive evaluation of the accuracy and reliability of our model.

Training Settings: We use RAFT and GMA as the backbones. During the pretraining stage, we add color augmentation and transformation augmentation to the pretraining dataset while for the finetuning stage, we follow the original setting of RAFT and GMA. In the pretraining stage, we calculate the farthest point distance and select 7,000 points.

Comparisons: We compare our pretraining task mainly with the two backbones, RAFT and GMA. For better comparison, we also use the given parameters for Sintel to finetune the SHOF. For the other methods, we follow [RRB18] and [RHT*20]. Besides, we also include several challenging baselines which are based on Transformer, including FlowFormer [HSZ*22], GMFlow [XZC*22]. In order to further evaluate the effectiveness of our pretraining method, we also include unsupervised loss functions introduced in [MHR18] as a baseline.

4.2. Quantitative Results

Table 1: *Results on SHOF. The **best** and the **second best** are highlighted correspondingly. The '-S' means using the given Sintel checkpoint and † means further finetune on SHOF.*

Method	SHOF	MHOF			
	AEPE ↓	AEPE ↓	1px ↑	3px ↑	5px ↑
HumanFlow [RRB18]	0.1164	/	/	/	/
PWC-Net [SYLK18]	0.2158	/	/	/	/
FlowFormer [HSZ*22]	0.0859	0.4742	0.9173	0.9744	0.9847
GMFlow [XZC*22]	**0.0817**	1.5049	0.5093	0.9168	0.9805
UnFlow-R [MHR18]	0.6790	1.2166	0.5581	0.9475	0.9740
RAFT-S† [TD20]	0.0912	0.4673	0.9173	**0.9756**	**0.9855**
RAFT	0.0873	0.4903	0.9156	0.9737	0.9845
Ours-RAFT-S†	0.0842	0.4675	**0.9186**	**0.9755**	**0.9853**
Ours-RAFT	**0.0761**	**0.4409**	**0.9193**	0.9753	0.9851
GMA-S† [JCL*21]	0.0938	**0.4570**	0.9120	0.9753	0.9851
GMA	0.0919	0.5427	0.9053	0.9722	0.9838
Ours-GMA	0.0874	0.4736	0.9162	0.9744	0.9845

We initially validate our quantitative results on the SHOF dataset, as illustrated in Tab. 1. Compared to RAFT and GMA models without pretraining, our method achieves improvements of 12.8% and 4.9% in AEPE on SHOF, respectively. For the relatively saturated task of SHOF, incorporating our pretraining still results in a noticeable improvement in performance. Additionally, the generalization performance on MHOF is enhanced by 10.1% and 30.9%, demonstrating the effectiveness of our approach in both specific and broader contexts. It is noteworthy that for GMA, our method significantly enhances the generalization effect on MHOF. This demonstrates that our pretraining approach aids the network in

Table 2: *Results on MHOF.The* ***best*** *and the* ***second best*** *are highlighted correspondingly. '-S' means using the given Sintel checkpoint and † means further finetune on MHOF.*

Method	MHOF			
	AEPE ↓	1px ↑	3px ↑	5px ↑
SPyNet† [RB17]	0.3910	/	/	/
PWC-Net† [SYLK18]	0.3010	/	/	/
FlowFormer [HSZ*22]	0.2246	0.9400	0.9880	0.9942
RAFT [TD20]	0.2338	0.9381	0.9850	0.9923
Ours-RAFT	**0.2194**	**0.9406**	**0.9882**	**0.9942**
GMA [JCL*21]	0.2253	0.9387	0.9874	0.9938
Ours-GMA	**0.2161**	**0.9408**	**0.9882**	**0.9943**

learning more fundamental features, rather than leading it to overfit a specific dataset. This aspect underscores the robustness and versatility of our pretraining strategy in improving model performance across different datasets. For UnFlow-RAFT (Tab.1 row 5), we pretrain on D-FAUST with the unsupervised losses from [MHR18] and then fine-tune on SHOF, all under the same setting as ours. The resulting AEPE on SHOF is 0.68, which is significantly larger than ours. [MHR18] assumes the input images are sufficiently similar to each other, which supports the smoothness and consistency loss. This, however, hinders it from modeling potential significant non-rigid deformations, which is the main goal of our method.

We then analyze the impact of pretraining dataset and initialization. First we train on dataset Sintel and then fine-tuning on SHOF for both backbones(RAFT in Tab.1 row 6 and GMA in row 10). The results of fine-tuning after training with strong label signals are not as effective as those achieved by fine-tuning after using our pretraining. Furthermore, we also pretrain on D-FAUST with initialization of Sintel checkpoint and then fine-tune on SHOF (Tab.1 row 8). The improvement turns out to be minor, which also highlights the efficiency of our designed framework. In addition, we compare our method with other optical flow methods. We use the published training details and train a new version for SHOF. For both performance accuracy and generalization ability, our methods indicate better performance. In addition, our method can achieve competing results with pretraining dataset of smaller size and does not introduce extra learning parameters for the backbone.

To assess the generalization capabilities of our method, we train on the MHOF dataset, with specific results shown in Tab. 2. Our pretraining is based on unsupervised training with the single-person flow. Despite this, both backbones show improved performance on MHOF after fine-tuning: we observed enhancements in AEPE by 6.2% and 4.1% on RAFT and GMA, respectively. This demonstrates that our pretraining method can effectively generalize to more complex, multi-person scenarios.

4.3. Qualitative Results

As is shown in Fig. 4, we first evaluate our pretrained model based on the RAFT network on the testing sequences of DeepDeform. We use the mask information to generate shape matching results for two frames and translate the per-pixel correspondence into optical flow. Even under large gaps between two frames, our model can

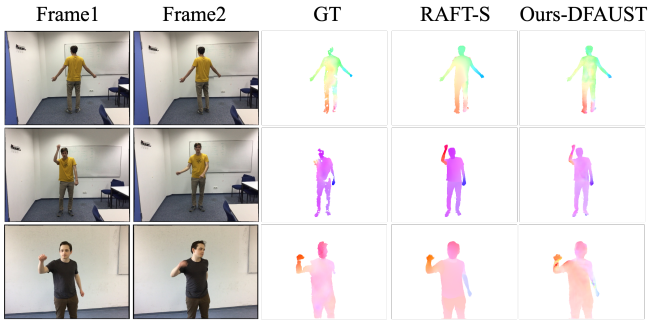

Figure 4: *Visualization of DeepDeform. RAFT-S refers to the RAFT using the given Sintel checkpoint. Ours-DFAUST refers to the optical flow transformed by the correspondence between the two non-rigid shapes which is only pretrained on D-FAUST.*

generate high-quality optical flow. Note that RAFT-S means RAFT using the given Sintel checkpoint which has been trained on five datasets, while our model is trained with unsupervised penalties on our synthetic pretraining dataset. The results demonstrate that our pretraining task can indeed help optical flow estimation.

We then visualize the results on SHOF. We mainly focus on the comparison with two backbones, RAFT and GMA. As is shown in Fig. 5, we calculate the per frame EPE to evaluate the results more precisely. Thanks to the transformation augmentation during the pretraining stage, the enhanced models can better separate the human motions from the background. Our pretraining method can also help fix some errors for human body.

For the results on MHOF, we also calculate the per-frame EPE. When dealing with the occlusion between different people, our pretrained models can better segment the individuals and generate precise optical flow for different individuals separately. For optical flow between two frames with the interval of 10, as is shown in Fig. 7, with our pretraining task, both RAFT and GMA can better handle large deformation human motions, such as bending over, raising legs, and turning around, indicating that our pretraining task can indeed guide the backbone and increase the accuracy of human optical flow estimation.

4.4. Ablation study

To validate the soundness of our pretraining framework design, we conduct ablation studies analyzing two crucial parameters in pretraining: the interval between selected frames and the inclusion of XY-axis translation data augmentation. The specific results are shown in Tab. 3. We choose RAFT as the backbone for training on the SHOF dataset and test it on both SHOF and MHOF datasets to better compare both the accuracy and generalization. Initially, we fix the frame selection interval at 20 frames. The addition of translation augmentation led to a 20% decrease in AEPE, which is logical given that the SHOF dataset includes flow translations. This result underscores the appropriateness of incorporating translation augmentation in our pretraining. Building on this foundation, we maintained consistent translation data augmentation in our experiments while varying the interval of frame selection. We observed that with fewer frames (5, 10), the pretraining process converged

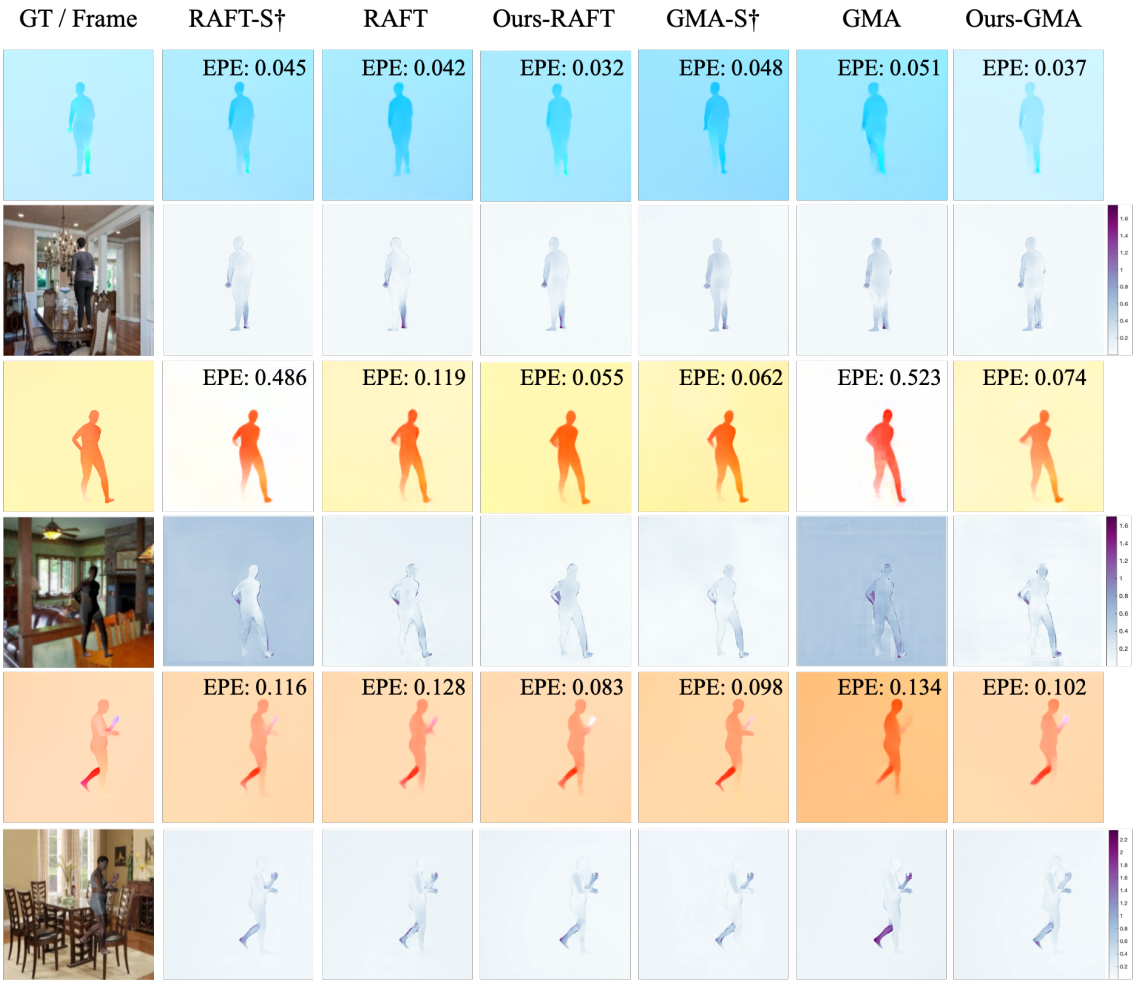

Figure 5: *Results on SHOF. Suffix '-S' means using the given Sintel checkpoint and '†' means further fine-tuning on SHOF. We demonstrate the estimated optical flow together with the regarding EPE (odd rows) and show the corresponding error maps (even rows).*

more quickly, but the fine-tuning results on SHOF and MHOF were sub-optimal. This outcome suggests that overly simplistic pretraining tasks offer limited benefits for downstream tasks. Conversely, excessively large frame intervals posed challenges in pretraining convergence, leading to poor generalization in downstream tasks. Consequently, we determined that an optimal frame interval for our pretraining is 20 frames.

Table 3: *Ablation results on SHOF under different settings.*

Ours-RAFT		SHOF	SHOF			
			MHOF			
Gap	Trans_aug	AEPE ↓	AEPE ↓	1px ↑	3px ↑	5px ↑
30	✓	0.0897	0.4872	0.9139	0.9750	0.9852
20	✓	0.0761	0.4409	0.9193	0.9753	0.9851
20	×	0.0835	0.4657	0.9171	0.9851	0.9856
10	✓	0.0848	0.4767	0.9159	0.9749	0.9851
5	✓	0.0897	0.4875	0.9125	0.9749	0.9851

5. Conclusion

In the hope of leveraging 3D geometric information to guide the 2D image feature, we design a novel pretraining task to enhance human optical flow estimation and our framework can be applied to most optical flow networks without additional learning parameters. We combine 2D pixel-wise features with 3D point-wise geometry through a mesh rendering component, which can be applied to other synthetic 3D shapes. By transforming the 2D optical flow into functional maps, we utilize the pure spectral features to enhance the optical flow estimation and indeed improve the performance of optical flow estimation for non-rigid situations. While our method performs less satisfactorily on details, such as hands. Thus, in the future, we will go on improving the performance.

Acknowledgement This work was supported by the National Natural Science Foundation of China under contract No. 62171256 and Meituan.

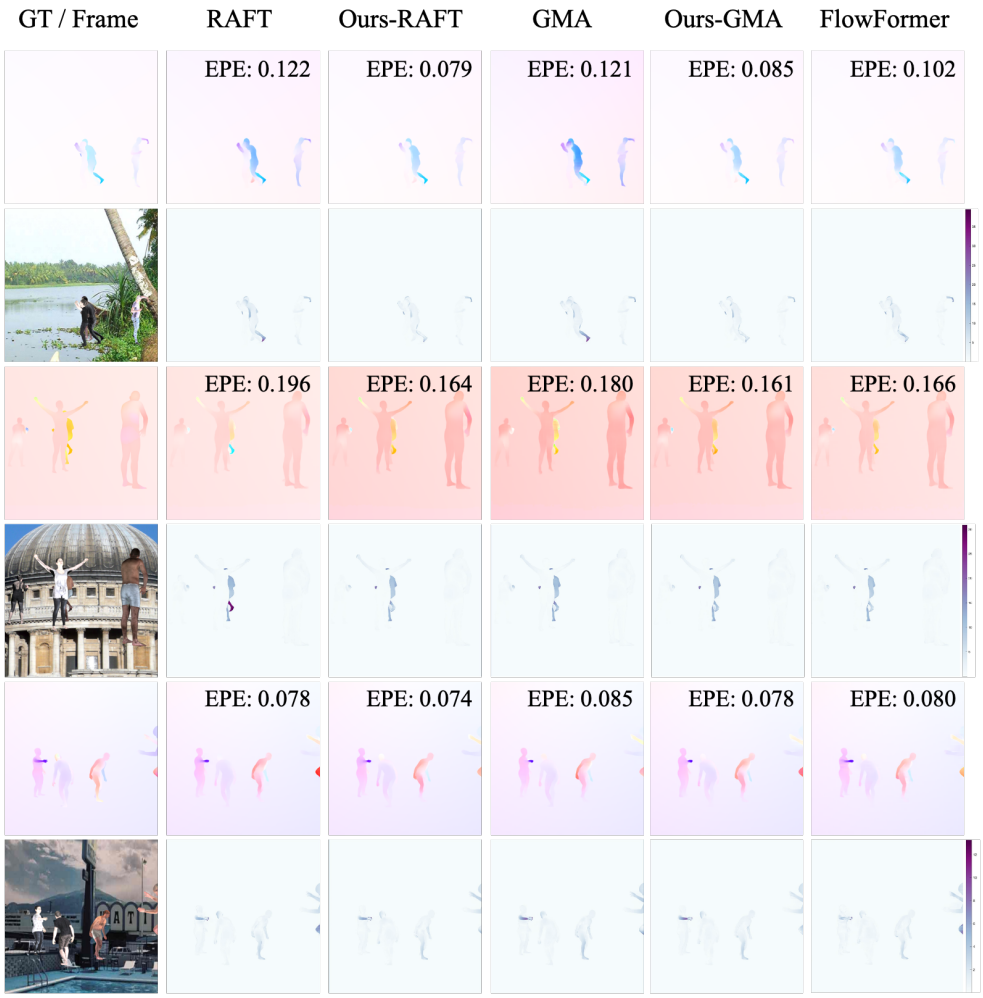

Figure 6: *Results on MHOF. We demonstrate the estimated optical flow together with the regarding EPE (odd rows) and show the corresponding error maps (even rows).*

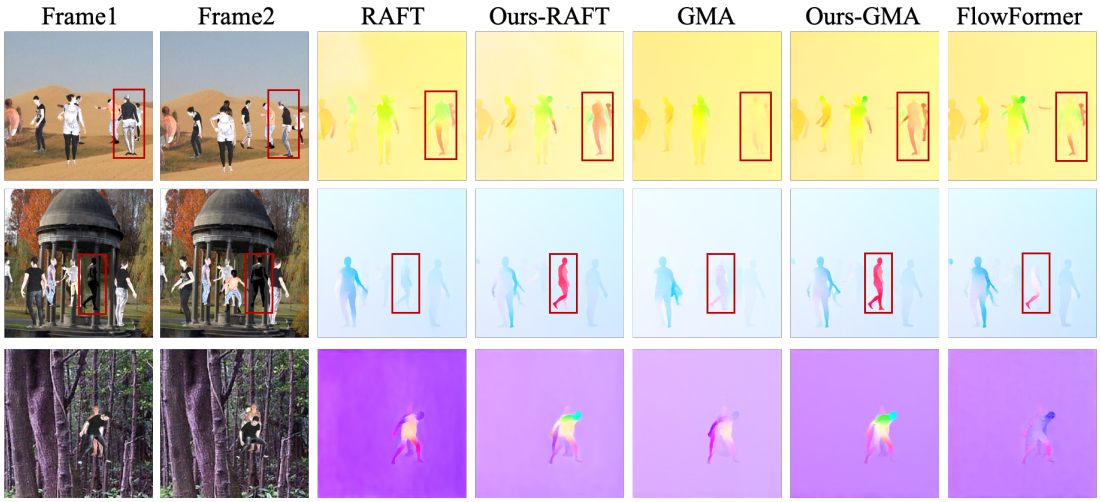

Figure 7: *Visualization of MHOF with the interval of 10 frames.*

References

[BA96] BLACK M. J., ANANDAN P.: The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer vision and image understanding 63*, 1 (1996), 75–104. 2

[BBPW04] BROX T., BRUHN A., PAPENBERG N., WEICKERT J.: High accuracy optical flow estimation based on a theory for warping. In *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part IV 8* (2004), Springer, pp. 25–36. 2

[BGSK22] BAI S., GENG Z., SAVANI Y., KOLTER J. Z.: Deep equilibrium optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 620–630. 2

[BRPMB17] BOGO F., ROMERO J., PONS-MOLL G., BLACK M. J.: Dynamic faust: Registering human bodies in motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 6233–6242. 3

[BWSB12] BUTLER D. J., WULFF J., STANLEY G. B., BLACK M. J.: A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12* (2012), Springer, pp. 611–625. 2

[BZTN20] BOZIC A., ZOLLHOFER M., THEOBALT C., NIESSNER M.: Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 7002–7012. 2, 5

[DCF23] DONG Q., CAO C., FU Y.: Rethinking optical flow from geometric matching consistent perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1337–1347. 1, 2

[DFI*15] DOSOVITSKIY A., FISCHER P., ILG E., HAUSSER P., HAZIRBAS C., GOLKOV V., VAN DER SMAGT P., CREMERS D., BROX T.: Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 2758–2766. 2

[ELC20] EISENBERGER M., LAHNER Z., CREMERS D.: Smooth shells: Multi-scale shape registration with functional maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 12265–12274. 3

[ERGB16] EYNARD D., RODOLA E., GLASHOFF K., BRONSTEIN M. M.: Coupled functional maps. In *2016 Fourth International Conference on 3D Vision (3DV)* (2016), IEEE, pp. 399–407. 5

[ETLTC20] EISENBERGER M., TOKER A., LEAL-TAIXÉ L., CREMERS D.: Deep shells: Unsupervised shape correspondence with optimal transport. *Advances in Neural information processing systems 33* (2020), 10491–10502. 3

[GLW*21] GOJCIC Z., LITANY O., WIESER A., GUIBAS L. J., BIRDAL T.: Weakly supervised learning of rigid 3d scene flow. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 5692–5703. 2

[HLR*19] HALIMI O., LITANY O., RODOLA E., BRONSTEIN A. M., KIMMEL R.: Unsupervised learning of dense shape correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 4370–4379. 3

[HR19] HUR J., ROTH S.: Iterative residual refinement for joint optical flow and occlusion estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 5754–5763. 2

[HRWO20] HUANG R., REN J., WONKA P., OVSJANIKOV M.: Consistent zoomout: Efficient spectral map synchronization. In *Computer Graphics Forum* (2020), vol. 39, Wiley Online Library, pp. 265–278. 3

[HS81] HORN B. K., SCHUNCK B. G.: Determining optical flow. *Artificial intelligence 17*, 1-3 (1981), 185–203. 2

[HSZ*22] HUANG Z., SHI X., ZHANG C., WANG Q., CHEUNG K. C., QIN H., DAI J., LI H.: Flowformer: A transformer architecture for optical flow. In *European Conference on Computer Vision* (2022), Springer, pp. 668–685. 1, 2, 5, 6

[IMS*17] ILG E., MAYER N., SAIKIA T., KEUPER M., DOSOVITSKIY A., BROX T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 2462–2470. 1, 2

[JCL*21] JIANG S., CAMPBELL D., LU Y., LI H., HARTLEY R.: Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 9772–9781. 2, 5, 6

[JSB*20] JONSCHKOWSKI R., STONE A., BARRON J. T., GORDON A., KONOLIGE K., ANGELOVA A.: What matters in unsupervised optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16* (2020), Springer, pp. 557–572. 2

[JSH23] JIANG P., SUN M., HUANG R.: Non-rigid shape registration via deep functional maps prior, 2023. arXiv:2311.04494. 3

[JYSP21] JI B., YANG C., SHUNYU Y., PAN Y.: Hpof: 3d human pose recovery from monocular video with optical flow. In *Proceedings of the 2021 International Conference on Multimedia Retrieval* (2021), pp. 144–154. 1

[KPD15] KALE K., PAWAR S., DHULEKAR P.: Moving object tracking using optical flow and motion vector estimation. In *2015 4th international conference on reliability, infocom technologies and optimization (ICRITO)(trends and future directions)* (2015), IEEE, pp. 1–6. 1

[LL21] LEE G.-H., LEE S.-W.: Uncertainty-aware human mesh recovery from video by learning part-based 3d dynamics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2021), pp. 12375–12384. 1

[LLX*22] LIU H., LU T., XU Y., LIU J., LI W., CHEN L.: Camliflow: Bidirectional camera-lidar fusion for joint optical flow and scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 5791–5801. 1, 2

[LMR*23] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 2023, pp. 851–866. 5

[LRR*17] LITANY O., REMEZ T., RODOLA E., BRONSTEIN A., BRONSTEIN M.: Deep functional maps: Structured prediction for dense shape correspondence. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 5659–5667. 3

[LTC*21] LEE Y.-C., TSENG K.-W., CHEN Y.-T., CHEN C.-C., CHEN C.-S., HUNG Y.-P.: 3d video stabilization with depth estimation by cnn-based optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 10621–10630. 1

[LXH*22] LI Z., XU B., HUANG H., LU C., GUO Y.: Deep two-stream video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (January 2022), pp. 430–439. 1

[LYL*22] LUO A., YANG F., LUO K., LI X., FAN H., LIU S.: Learning optical flow with adaptive graph reasoning. In *Proceedings of the AAAI conference on artificial intelligence* (2022), vol. 36, pp. 1890–1898. 1, 2

[LYL*23] LUO A., YANG F., LI X., NIE L., LIN C., FAN H., LIU S.: Gaflow: Incorporating gaussian attention into optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 9642–9651. 2

[LZL*22] LI R., ZHANG C., LIN G., WANG Z., SHEN C.: Rigidflow: Self-supervised scene flow learning on point clouds by local rigidity prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 16959–16968. 1, 2

[LZYX22] LIN W., ZHENG C., YONG J.-H., XU F.: Occlusionfusion:

Occlusion-aware motion estimation for real-time dynamic 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 1736–1745. 1

[MHR18] MEISTER S., HUR J., ROTH S.: Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI conference on artificial intelligence* (2018), vol. 32. 2, 5, 6

[MIH*16] MAYER N., ILG E., HAUSSER P., FISCHER P., CREMERS D., DOSOVITSKIY A., BROX T.: 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* (2016), pp. 4040–4048. 2

[MRR*19] MELZI S., REN J., RODOLA E., SHARMA A., WONKA P., OVSJANIKOV M.: Zoomout: Spectral upsampling for efficient shape correspondence. *arXiv preprint arXiv:1904.07865* (2019). 3

[OBCS*12] OVSJANIKOV M., BEN-CHEN M., SOLOMON J., BUTSCHER A., GUIBAS L.: Functional maps: a flexible representation of maps between shapes. *ACM Transactions on Graphics (ToG) 31*, 4 (2012), 1–11. 1, 3, 4, 5

[RB17] RANJAN A., BLACK M. J.: Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 4161–4170. 2, 6

[RHT*20] RANJAN A., HOFFMANN D. T., TZIONAS D., TANG S., ROMERO J., BLACK M. J.: Learning multi-human optical flow. *International Journal of Computer Vision 128* (2020), 873–890. 2, 5

[ROA*13] RUSTAMOV R. M., OVSJANIKOV M., AZENCOT O., BEN-CHEN M., CHAZAL F., GUIBAS L.: Map-based exploration of intrinsic shape differences and variability. *ACM Transactions on Graphics (TOG) 32*, 4 (2013), 1–12. 5

[RRB18] RANJAN A., ROMERO J., BLACK M. J.: Learning human optical flow. *arXiv preprint arXiv:1806.05666* (2018). 2, 5

[RSO19] ROUFOSSE J.-M., SHARMA A., OVSJANIKOV M.: Unsupervised deep learning for structured shape matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 1617–1627. 2, 3, 4

[RTB22] ROMERO J., TZIONAS D., BLACK M. J.: Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610* (2022). 5

[SBL*23] SUN Y., BAO Q., LIU W., MEI T., BLACK M. J.: Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2023), pp. 8856–8866. 1

[SCZ*22] SUN S., CHEN Y., ZHU Y., GUO G., LI G.: Skflow: Learning optical flow with super kernels. *Advances in Neural Information Processing Systems 35* (2022), 11313–11326. 2

[SHB*23] SHI X., HUANG Z., BIAN W., LI D., ZHANG M., CHEUNG K. C., SEE S., QIN H., DAI J., LI H.: Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. *arXiv preprint arXiv:2303.08340* (2023). 1, 2

[SHL*23] SHI X., HUANG Z., LI D., ZHANG M., CHEUNG K. C., SEE S., QIN H., DAI J., LI H.: Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1599–1610. 1, 2

[SMJ*23] SUN M., MAO S., JIANG P., OVSJANIKOV M., HUANG R.: Spatially and spectrally consistent deep functional maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 14497–14507. 3, 4

[SYLK18] SUN D., YANG X., LIU M.-Y., KAUTZ J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 8934–8943. 1, 2, 5, 6

[TD20] TEED Z., DENG J.: Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16* (2020), Springer, pp. 402–419. 1, 2, 5, 6

[TD21] TEED Z., DENG J.: Raft-3d: Scene flow using rigid-motion embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 8375–8384. 1, 2

[WLL*22] WEINZAEPFEL P., LEROY V., LUCAS T., BRÉGIER R., CABON Y., ARORA V., ANTSFELD L., CHIDLOVSKII B., CSURKA G., REVAUD J.: Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems 35* (2022), 3502–3516. 1, 2

[WLL*23a] WEINZAEPFEL P., LUCAS T., LEROY V., CABON Y., ARORA V., BRÉGIER R., CSURKA G., ANTSFELD L., CHIDLOVSKII B., REVAUD J.: Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 17969–17980. 1, 2

[WLL*23b] WU G., LIU X., LUO K., LIU X., ZHENG Q., LIU S., JIANG X., ZHAI G., WANG W.: Accflow: Backward accumulation for long-range optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 12119–12128. 2

[XZC*22] XU H., ZHANG J., CAI J., REZATOFIGHI H., TAO D.: Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 8121–8130. 2, 5

[YHD16] YU J. J., HARLEY A. W., DERPANIS K. G.: Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14* (2016), Springer, pp. 3–10. 2

[ZSD*20] ZHAO S., SHENG Y., DONG Y., CHANG E. I., XU Y., ET AL.: Maskflownet: Asymmetric feature matching with learnable occlusion mask. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 6278–6287. 2