METACAP: Meta-learning Priors from Multi-View Imagery for Sparse-view Human Performance Capture and Rendering

Guoxing Sun[1], Rishabh Dabral[1], Pascal Fua[2], Christian Theobalt[1], Marc Habermann[1]

[1] Max Planck Institute for Informatics, Saarland Informatics Campus
[2] EPFL
{gsun,rdabral,theobalt,mhaberma}@mpi-inf.mpg.de, pascal.fua@epfl.ch
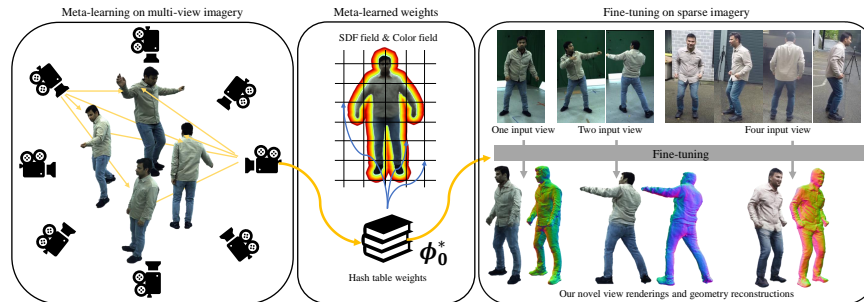https://vcai.mpi-inf.mpg.de/projects/MetaCap/

Fig. 1: We propose METACAP, a new approach for capturing 3D humans from sparse-view or even monocular images, which enables high fidelity 3D geometry recovery and photoreal free-view synthesis within a range of seconds to minutes.

Abstract. Faithful human performance capture and free-view rendering from sparse RGB observations is a long-standing problem in Vision and Graphics. The main challenges are the lack of observations and the inherent ambiguities of the setting, e.g. occlusions and depth ambiguity. As a result, radiance fields, which have shown great promise in capturing high-frequency appearance and geometry details in dense setups, perform poorly when naïvely supervising them on sparse camera views, as the field simply overfits to the sparse-view inputs. To address this, we propose METACAP, a method for efficient and high-quality geometry recovery and novel view synthesis given very sparse or even a single view of the human. Our key idea is to meta-learn the radiance field weights solely from potentially sparse multi-view videos, which can serve as a prior when fine-tuning them on sparse imagery depicting the human. This prior provides a good network weight initialization, thereby effectively addressing ambiguities in sparse-view capture. Due to the articulated structure of the human body and motion-induced surface deformations, learning such a prior is non-trivial. Therefore, we propose to meta-learn the field weights in a pose-canonicalized space, which reduces the spatial feature range and makes feature learning more effective. Consequently, one can fine-tune our field parameters to quickly generalize to

unseen poses, novel illumination conditions as well as novel and sparse (even monocular) camera views. For evaluating our method under different scenarios, we collect a new dataset, WILDDYNACAP, which contains subjects captured in, both, a dense camera dome and in-the-wild sparse camera rigs, and demonstrate superior results compared to recent state-of-the-art methods on, both, public and WILDDYNACAP dataset.

Keywords: Human Performance Capture · Meta Learning

1   Introduction

Human performance capture has made enormous strides in recent years. Typically, the quality of the reconstruction improves with the number of cameras being used. The highest-quality results are obtained using many—8 or more—calibrated cameras in a dome-like configuration [9, 17, 64, 67, 92]. However, such capture domes are expensive and difficult to set up, which restricts their applicability. A truly practical solution ought to be easy to deploy while supporting arbitrary camera configurations, including *sparse* ones, i.e. four or less cameras. Unfortunately, sparsity comes with its own challenges, such as self-occlusions and depth ambiguities, which often result in lower quality outcomes.

How to mitigate the effects of these ambiguities and missing information is the core problem of sparse-view reconstructions and researchers have explored different priors to compensate for the ambiguities of such an ill-posed problem. Earlier works introduce priors on template meshes [20, 21, 27, 85] to solve the above ambiguities. However, meshes suffer from limited resolution and are not easy to incorporate into learning frameworks. Implicit surface representations offer a promising alternative and some methods [59, 94] can learn implicit field priors from large-scale human scan datasets. However, they suffer from blurry appearance and poor generalization capability because the amount of data is limited and model capacity is small. While many works have recently focused on animatable implicit radiance fields for humans [18, 37, 55, 78, 80, 93], only few [16, 38, 48, 87] have explored priors for optimizing or fitting a neural implicit field given sparse imagery. They usually only show results for *simple objects*, but not for complex structures like articulated humans.

Recently, some works [68, 72] showed that meta-learning can learn powerful neural field priors for few-shot learning. To learn the field prior for personalized high fidelity human capture more effectively, we argue that we should decompose the human model into a coarse-level geometry represented by mesh-based templates [19, 39] and a fine-level neural implicit field [76, 79, 92] accounting for details such as clothing wrinkles. The coarse-level prior from multi-view images is easier to obtain and can serve as a crucial component in fine-level prior learning. At the fine level, we, for the first time in literature, *explore learning a meta prior from images for an implicit human representation*, which at test time can be optimized using sparse image cues, thereby allowing fast adaptation to unseen poses, novel views, different camera setups, and novel illumination conditions.

Our method, dubbed METACAP, represents the fine-level human body as a signed distance field (SDF) and an appearance field parameterized with an efficient hashgrid encoding [45], which can be volume rendered [76] into an image. To learn an effective prior for such a representation, we propose to meta-learn the hashgrid parameters solely from multi-view imagery during training in an end-to-end manner, i.e. learning the optimal weights, which, at test time under the sparse settings, yield improved convergence rate and accuracy. However, naïvely meta-learning these implicit parameters for static scenes [68,72], leads to poor performance on complicated settings such as human performance capture (Fig. 5). This is because the human surface is highly articulated and always deforms. Thus, learning hashgrid parameters, which live on a spatially fixed grid, is not effective, e.g. the limbs can be mapped onto entirely different hash table parameters depending on their articulation. Therefore, we further propose a space canonicalization step, i.e. we transform points from global space to a canonical pose space. To obtain the transform, we query the nearest transformation of a coarse-level human template. We highlight that this canonicalization is not tightly bound to a specific template, but supports deformable human models [19] as well as parametric (piece-wise rigid) body models [39]. Additionally, we introduce occlusion handling, where we use a visibility map to guide the ray sampling, enabling reconstruction of heavily occluded regions at sparse or even monocular inputs at inference time.

We evaluate our method on the DynaCap [19] benchmark, which provides multi-view recordings of humans in a controlled studio setup. However, DynaCap lacks in-the-wild scenes for evaluating the robustness to various environment conditions such as the difference in lighting and camera parameters. Thus, we provide a new dataset, WILDDYNACAP, containing paired *dense multi-view studio* captures and *sparse multi-view in-the-wild* recordings for each subject. In summary, our main contributions are:

- A meta-learning method learned from multi-view imagery for high-quality reconstruction of human geometry and appearance under sparse cameras.
- At the technical core, we propose a meta-learning strategy to learn the optimal weights of an implicit human representation solely from multi-view images, which effectively serves as a prior when deployed to the sparse reconstruction task.
- We further demonstrate the importance of space canonicalization for the human-specific meta-learning task, and introduce a dedicated occlusion handling strategy.

Our quantitative and qualitative experiments demonstrate that METACAP achieves state-of-the-art geometry recovery and novel view synthesis compared to prior works. Our evaluations also demonstrate that METACAP generalizes to novel poses and induced surface deformations, change of lighting conditions, and camera parameters. Moreover, we highlight the versatility of our approach as our meta-learning strategy can be supervised on an arbitrary number of views also including monocular videos. Similarly, during fine-tuning, our approach supports reconstruction from sparse in-the-wild multi-view images as well as monocular

imagery (see Fig. 1). Last, our method is agnostic to the choice of the deformable human model, thus, even supporting loose types of apparel.

2   Related Work

In the following, we discuss prior works on (1) scene-agnostic implicit representations for novel view synthesis and reconstruction, (2) human-specific sparse-view reconstruction and rendering methods, (3) personalized performance capture and (4) 3D reconstruction using meta-learning. Animatable avatar methods [18, 19, 32, 34, 37, 40, 41, 51, 61, 65, 95, 98], which only take the skeletal pose as input during inference, are not in our scope, since we focus on the reconstruction task, i.e. recovering geometry and appearance from 2D imagery during inference.
Scene-agnostic Implicit Representations. The emergence of implicit representations for 3D scene reconstruction [42, 52, 59, 76] and novel-view synthesis [3–5, 44] paved the way towards compact and high-fidelity scene representations. Subsequent methods have attempted to improve upon a variety of aspects, such as faster training [7, 13, 45], faster inference [29, 88], sparse [8, 16, 38] and monocular-view [15, 89] reconstruction. In our work, we build upon the scene representation introduced in NeuS2 [79] and Instant-NSR [92]. They leverage multi-grid hash encoding [45] for parameterizing the implicit surface [76] to enable fast and high quality reconstruction. However, they require sufficiently dense multi-view images to ensure accurate geometry and appearance recovery. Due to occlusions and the depth ambiguity inherently present in the sparse setup, simply optimizing such methods with sparse RGB imagery leads to poor reconstructions. More recently, some methods explored different geometry priors [16, 38] or regularizations [28, 48, 87] to enable few-shot novel view synthesis. However, none of them proves to be effective under large camera baselines or sparse 360° scenarios, where the image observations are very limited and occlusions happen more frequently. We address these challenges by meta-learning an optimal set of initial hashgrid parameters, which effectively serves as a prior during inference on sparse observations, and a human-specific space canonicalization.
Sparse-View Performance Capture and Rendering. Beyond occlusions and the depth ambiguity, sparse performance capture comes with the additional challenges of modeling and capturing the body articulations as well as complex surface deformations of the clothes. To compensate for the absence of dense supervision, sparse human reconstruction methods typically employ additional priors. They range from using a parametric template for a coarse initialization of the geometry [54–56, 62, 94, 96], to using depth supervision [49, 70, 86, 90, 99] and data-driven priors [10, 24, 59, 60, 64, 73, 77, 83, 84, 94, 97]. Several volumetric rendering-based methods focus on learning the human geometry and appearance in the canonical space [26, 50, 54, 56, 58, 63, 74, 75], which learns shared features across different poses. While such methods excel in novel view synthesis tasks, the recovered geometry is often coarse. Some works [6, 31, 43, 50] learn radiance field priors from large scale datasets. However, their conventional learning strategy limits their fine-tuning ability. Another key challenge in human reconstruction is

associated with the modeling of loose clothing. Previous methods have attempted to address this by optimizing the vertex deformations on top of a template model [19–21, 41, 82] or by modeling clothes as a separate layer [11, 75]. In this work, we utilize a hybrid representation where we leverage an explicit human model to facilitate meta-learning an optimal set of weights parameterizing an implicit field in a pose-canonicalized space.

Personalized Performance Capture. In contrast to generalizable performance capture methods [50, 59, 78, 94], personalized human capture first builds a person-specific prior on multi-view data [10, 21, 35, 36, 66, 74, 81] while during inference the avatar can be reconstructed from sparser signals, e.g. a monocular image. Therefore, the goal of such methods is less about cross-identity generalization, but rather highest-fidelity surface and appearance recovery from sparse sensory data while ensuring generalization to different human motions, lighting condition, camera poses, camera numbers, and surface dynamics. Our setting is most closely related to these methods. However, we for the first time in literature propose meta-learning as a prior on the implicit human representation solely from images while experimentally showing that we outperform previous methods operating under the same setting.

Reconstruction with Meta-Learning. Meta-learning is intended to learn from multiple tasks or a single task [23]. When new observations are presented, domain adaptation or improved performance can be achieved with minimal training iterations. Here, we focus on optimization-based methods [2, 12, 33, 47, 57], especially MAML [12] and Reptile [47]. MetaSDF [68] learns SDF priors and achieves faster inference than auto-decoder methods [52]. Tancik et al. [72] apply a similar strategy to different coordinate-based neural representations, such as image fields and radiance fields [44]. MetaAvatar [77] extends this idea from static to dynamic human representations. After meta-learning on 3D human scans [41], the learned SDF weights can be fine-tuned using a few depth maps.

ARAH [78] is methodically most closely related to our work. However, there are significant differences: 1) They meta-learn human priors from 3D scans while only supporting image-only supervision during fine-tuning. In stark contrast, our formulation allows meta-learning from potentially sparse images directly as we deeply entangle the volume rendering and hash-grid parameterization with the meta-learning routine. Our end-to-end prior learning and fine-tuning, thus, enables better recovery of high-frequency geometry and appearance. 2) Our whole design, i.e. space canonicalization and scene parameterization, is geared towards efficiency, significantly improving the fine-tuning time, reducing from hours (ARAH) to minutes (METACAP). 3) We experimentally show that our canonicalization is agnostic to specific template choices, thus, we can also account for loose clothing while ARAH solely shows tightly clothed people.

3   Method

Our goal is to recover high-quality human models from few, or even only one, RGB image. Let us consider an implicit function f_ϕ parameterized by a network
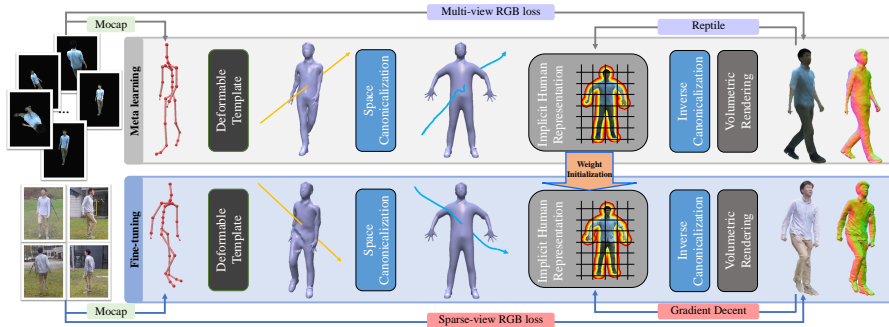
Fig. 2: Overview. METACAP is a novel performance capture method, which meta-learns the pose-canonical and optimal network weights of an implicit human representation solely from multi-view images. At inference, with only sparse in-the-wild images, those weights effectively act as a prior and resolve the inherent ambiguities, i.e. occlusion and depth, while maintaining high geometric and visual quality.

with weights ϕ that encodes the geometry and the appearance of the clothed human. Our sparse-view reconstruction method comprises two phases, multi-view meta-learning followed by sparse-view fine-tuning, as shown in Fig. 2.

During the meta-learning phase (Sec. 3.1 and 3.2), we are given a multi-view video using K_d cameras, which captures the performer. We denote the images of frame f as $\mathcal{I}_f = \{I_f^k | k \in [1; K_d]\}$ and assume the availability of paired masks $\mathcal{O}_f$ and the corresponding skeletal motion $\mathcal{M}_{f,W} = \{\boldsymbol{\theta}_{f-W}, \boldsymbol{\alpha}_{f-W}, \boldsymbol{z}_{f-W}, ..., \boldsymbol{\theta}_f, \boldsymbol{\alpha}_f, \boldsymbol{z}_f\}$, defined over a window of W frames, with $\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{z}$ representing the joint rotations, the root rotation, and the root translation, respectively. Given this, we aim at meta-learning a personalized and optimal set of weights ϕ^* for the implicit human representation, effectively serving as a data-driven prior.

During the fine-tuning phase (Sec. 3.3), we fine-tune the meta-learned weights ϕ^* to sparse in-the-wild images $\mathcal{I}_f = \{I_f^k | k \in [1; K_s]\}$ paired with masks $\mathcal{O}_f = \{O_f^k | k \in [1; K_s]\}$ with $K_s << K_d$. Notably, the motion $\mathcal{M}_{f,W}$ and surface dynamics of the performer as well as the lighting conditions and camera configuration during fine-tuning can significantly differ from the ones observed during training. We later show that fine-tuning the meta-learned weights ϕ^* on $\mathcal{I}_f$ leads to faster and better convergence compared to random weights ϕ_0.

3.1   Meta-Learning a Reconstruction Prior

Meta-Learning [12, 47] trains a neural model in a way that makes it easier to adapt to novel tasks. In our case, the novel tasks are sparse-view reconstruction and novel-view synthesis under unseen motions as well as lighting conditions. For meta-learning the 3D geometry and appearance of the human, we represent the human as an implicit SDF and radiance field. Given a ray $\mathbf{r}(\mathbf{o}, \mathbf{v})$ emanating from an image $\mathcal{I}_f$ with origin $\mathbf{o}$, and viewing direction $\mathbf{v}$, we sample n points $\{\mathbf{p}_i = \mathbf{o} + t_i \mathbf{v} | i \in [1; n]\}$ on depths t_i, and the implicit network f_ϕ predicts

the color $\mathbf{c}_i$ and the SDF s_i for each sample. We then utilize unbiased volume rendering [76] along ray $\mathbf{r}$ to obtain RGB value $\hat{C}(\mathbf{r})$ and mask value $\hat{M}(\mathbf{r})$:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{n} T_i \alpha_i \mathbf{c}_i, \quad \hat{M}(\mathbf{r}) = \sum_{i=1}^{n} T_i \alpha_i, \quad \alpha_i = \max\big(\frac{\Psi(s_i) - \Psi(s_{i+1})}{\Psi(s_i)}, 0\big), \quad (1)$$

where T_i is the accumulated transmittance, α_i is the unbiased weight function, and $\Psi(\cdot)$ represents the derivative of a sigmoid function. These rendered rays can be used to supervise the implicit network $f_\phi(\cdot)$ using the following loss:

$$\mathcal{L}(f_\phi, \mathbf{R}) = \lambda_c \mathcal{L}_{\text{color}} + \lambda_e \mathcal{L}_{\text{eik}} + \lambda_m \mathcal{L}_{\text{mask}} + \lambda_s \mathcal{L}_{\text{sparse}}, \quad (2)$$

where $\mathbf{R} = \{\mathbf{r}_k | k \in [1; m]\}$ is a collection of m rays, $\mathcal{L}_{\text{color}} = \frac{1}{m} \sum_k \mathcal{H}(\hat{C}(\mathbf{r}_k), C(\mathbf{r}_k))$ is the Huber loss [25], $\mathcal{L}_{\text{eik}} = \frac{1}{mn} \sum_{k,i} \big| 1 - \langle \mathbf{n}(\mathbf{p}_{k,i}), \nabla_{\mathbf{p}_{k,i}} f_\phi(\mathbf{p}_{k,i}) \rangle \big|_2^2$ is the Eikonal loss [14] on the normal $\mathbf{n}(\mathbf{p}_{k,i})$ of the sample point $\mathbf{p}_{k,i}$. $\mathcal{L}_{\text{mask}} = \frac{1}{m} \sum_k BCE(\hat{M}(\mathbf{r}_k), O(\mathbf{r}_k))$ is the binary cross entropy loss. $s(\mathbf{p}_{k,i})$ is the SDF value of point $\mathbf{p}_{k,i}$, and $\mathcal{L}_{\text{sparse}} = \frac{1}{mn} \sum_{k,i} \exp^{-|s(\mathbf{p}_{k,i})|}$ is a sparseness regularization term [38]. $C(\mathbf{r}_k)$ and $O(\mathbf{r}_k)$ are the ground truth color and mask values of ray $\mathbf{r}_k$. To accelerate the rendering process, we parameterize the implicit field using a multi-resolution hashgrid [45]. Thus, the overall learnable parameters ϕ include the MLP weights of the implicit function as well as the hashgrid parameters.

   Now, the initial weights ϕ_0 for our coordinate-based implicit function f_ϕ can be learned through optimization-based meta-learning algorithms, like MAML [12] or Reptile [47]. Let us assume we have a dataset of tasks following the distribution $\mathcal{T}$, where a task $T \sim \mathcal{T}$ is defined as a set $\{\mathcal{L}, \{\mathbf{R}_f \sim \mathcal{I}_f, \mathcal{O}_f\}, f_\phi\}_f$ comprising the loss function, the input rays and the implicit function f_ϕ. Our meta-learning optimization is performed in two nested loops. For the jth iteration of the inner loop with sampled rays $\mathbf{R}^j$, it follows a gradient-descent optimization of the model parameters ϕ_j' using a learning rate l_{in}:

$$\phi_j' \leftarrow \phi_{j-1}' - l_{\text{in}} \nabla_\phi \mathcal{L}(f_\phi, \mathbf{R}^j)|_{\phi = \phi_{j-1}'}. \quad (3)$$

We leverage the model-agnostic meta-learning algorithm, Reptile [47], to optimize the initial weights by nesting an outer loop on top of the inner loop. The outer loop then updates the meta-learned parameters ϕ_i through the first-order update equation:

$$\phi_i \leftarrow \phi_{i-1} + l_{\text{out}}(\phi_M'(\phi_{i-1}, T_i) - \phi_{i-1}), \quad (4)$$

where T_i are the tasks sampled at the ith outer iteration, l_{out} denotes the learning rate of outer loop. $\phi_M'(\phi_{i-1}, T_i)$ represents the network weights initialized by the inner-loop with ϕ_{i-1}, which are then optimized for M steps on T_i. The resulting weights, ϕ_N, at the end of N outer loop iterations serve as the optimal initial weights ϕ^* for the fine-tuning stage. To keep the meta-learning computation tractable, we choose Reptile [47] in Eq. 4 as it uses a first-order approximation instead of MAML's second order gradient update.

   Note that the above formulation directly meta-learns a volumetric rendering-based implicit network in the world-space. This proves to be suboptimal (as

shown in Fig. 5 and Tab. 3) as humans are highly articulated; a challenge further exacerbated by loose clothing. In the task set T consisting of humans in different body poses, the meta-learned hash encoding can map the same body point into different hash-table parameters. To address this issue, we modify the volume rendering formulation in Eq. 1 by proposing our human template guidance.

3.2   Template-guided Meta-Learning

Human Template. Here, we introduce a generic motion-driven human template, defined as a deformable mesh $\bar{\mathbf{X}} \in \mathbb{R}^{V \times 3}$, of a character in the canonical pose $\bar{\mathcal{M}}$ (as shown in Fig. 2). The template mesh $\bar{\mathbf{X}}$ can be deformed according to a given skeletal motion $\mathcal{M}_{f,W}$. Each vertex in the human template first undergoes a per-vertex motion-dependent deformation in the canonical space through a transform $\mathbf{T}_{\mathrm{def}}(\mathcal{M}_{f,W}) \in \mathbb{R}^{4 \times 4}$. After being deformed locally with $\mathbf{T}_{\mathrm{def}}$, one can perform forward kinematics $\mathbf{T}_{\mathrm{FK}}(\mathcal{M}_f, \bar{\mathcal{M}})$ to transform the deformed canonical space vertices $\mathbf{T}_{\mathrm{def}}\bar{\mathbf{X}}$ to world space vertices $\mathbf{X}$. The total invertible transformation from the template's canonical space to world space is: $\mathbf{T} = \mathbf{T}_{\mathrm{FK}}\mathbf{T}_{\mathrm{def}}$ and the template vertices are transformed as $\mathbf{X} = \mathbf{T}\bar{\mathbf{X}}$.

Template-guided Ray Warping. The template mesh defined above provides us a coarse geometry information, which effectively guides our meta-learning since we are now able to learn the meta-weights in a pose-canonical space. The sampled points $\mathbf{p}_i$ can be warped from the world space into the canonical space through our proposed template guided ray warping. Concretely, we first project each sampled point onto the nearest face (triangle) of the posed template mesh $\mathbf{X}$. To compute the transformation matrix $\mathbf{T}_i$, we barycentricaly interpolate the transformation of the vertices defining the nearest triangle. Next, we apply the inverse transformation $\mathbf{T}_i^{-1}$ of the sampled point to obtain its canonical position $\bar{\mathbf{p}}_i = \mathbf{T}_i^{-1}\mathbf{p}_i$. The canonical points, $\bar{\mathbf{p}}_i$, can now be encoded with the learnable hash encoding and rendered using the volumetric rendering approach described in Eq. 1. Moreover the template allows us to perform empty skipping [22] for point samples whose distance to the template exceeds a threshold η, thereby significantly accelerating the convergence. In this way, our final meta-learning approach (see also Alg. 1) is more efficient and effective compared to the baseline, i.e. learning in global space.

3.3   Occlusion Handling and Fine-tuning

At inference, when only the calibrated monocular or sparse-view RGB images $\mathcal{I}_f$, masks $\mathcal{O}_f$, and skeletal motion $\mathcal{M}_{f,W}$ are available, we initialize our implicit human field with the meta-learned weights ϕ^*, conduct the space canonicalization with the fitted template $\mathbf{X}$ (Sec.3.2), and fine-tune using the loss from Eq.2. This can also be denoted as the task $\{\mathcal{L}, \{\mathbf{R}_f \sim \mathcal{I}_f, \mathcal{O}_f\}, f_\phi | \phi_0 = \phi^*\}_f$.

Occlusion Handling. For extreme occlusion cases, which may occur in the monocular setting (see also Fig. 3), we further propose a dedicated occlusion handling strategy by leveraging the posed human template. Inspired by [71], we first pre-build an implicit human field on one frame of dense images $\mathcal{I}_f$

Algorithm 1 METACAP's Meta-Learning Procedure

Initialize: Weights ϕ, outer/inner learning rate l^{out} and l^{in}
for $i = 1, \ldots, N$ **do**
    Sample frames $\{f^{(j)}\}_{j=1}^M$, views $\{k^{(j)}\}_{j=1}^M$ and generate rays $\{\mathbf{R}^{(j)}\}_{j=1}^M$
    Set $\phi'_0 = \phi$
    for $j = 1, \ldots, M$ **do**
        Find closest triangle of points $\mathbf{p}$ and compute the point to face distance d
        Filter points with d larger than threshold η
        Compute inverse transformation $\mathbf{T}$
        Transform points into canonical space $\bar{\mathbf{p}} = \mathbf{T}^{-1}\mathbf{p}$
        $\phi'_j = \phi'_{j-1} - l^{in}\nabla_\phi \mathcal{L}(f_\phi(\bar{\mathbf{R}}^{(j)})|_{\phi=\phi'_{j-1}})$
    end for
    $\phi \leftarrow \phi + l^{out}(\phi'_M - \phi)$
end for
Result: Optimal weights $\phi^* \leftarrow \phi$

Monocular Input   Results without OH   Visibility Map and Proxy Image   Results with OH
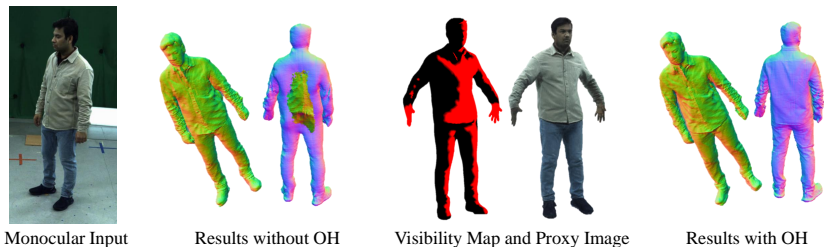
Fig. 3: Occlusion Handling. Our occlusion handling (abbreviated as OH) helps to regularize largely occluded regions.

or multiple frames of (in-the-wild) sparse images $\mathcal{I}_f$, and then render virtual views of the canonical space as proxy RGB images $\mathcal{I}^{\mathrm{proxy}}$. Next, during the fine-tuning stage, we determine per-vertex visibility by checking whether the vertex of the template is visible in any of the sparse input views. Finally, in addition to sampling rays from the input images, we sample additional virtual rays for the occluded regions from $\mathcal{I}^{\mathrm{proxy}}$. In practice, we found that the occlusion handling is only required in the case of monocular reconstruction (also see ablations).

4   Results

Implementation Details. For meta-learning, we perform $M = 24$ inner-loop gradient steps with $l^{out} = 1.0$ and $l^{in} = 1e - 4$. The inner loop optimizer is Adam [30] while the outer one uses stochastic gradient descent. Following NeuS [76], we perform a warmup in the inner loop at the beginning of training. If not stated otherwise, we use DDC [19] for space canonicalization, $\sim$80 cameras for meta learning, 4 cameras for fine-tuning in comparisons and ablations. For more details, please refer to the supplemental material.
Datasets.   We evaluate our approach on the publicly available dense-view
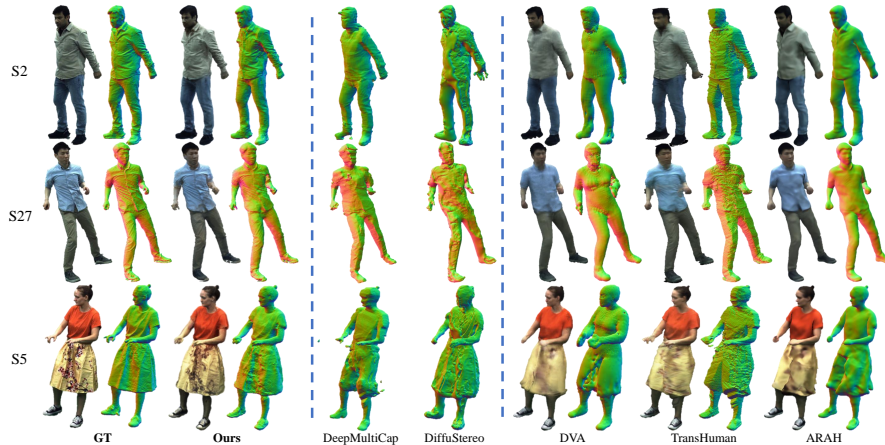
Fig. 4: Qualitative Comparison. Compared to previous works, our method better captures high-frequency geometry on cloth wrinkles and faithfully recovers appearance details such as facial details and clothing textures.

dome dataset DynaCap [19] and our WILDDYNACAP dataset. DynaCap provides around 100-view videos, foreground masks, and motions. We choose two subjects wearing loose (S5) and tight clothing (S3 in the supplemental). To compensate for the lack of in-the-wild sequences in DynaCap, we collect a new dataset WILDDYNACAP. It contains 110 cameras to capture the performer inside a dome, and a movable 5-camera setup to record videos in different in-the-wild scenarios. We captured two subjects in two camera setups. We recover the skeletal motion using markerless motion capture [69] using 34 views for the dome setups and 5 views for the in-the-wild sequences. For all experiments and methods, we use this motion (if not stated otherwise) as this work focuses on surface recovery rather than motion capture. We also provide ablation on motions only from sparse cameras in the supplemental.

Metrics and Ground Truth. To obtain the ground truth geometry for the dome captures, we leverage the recent state-of-the-art implicit reconstruction method InstantNSR [92] trained on the *dense* RGB views. To evaluate the rendering quality, we report the peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS) [91]. To evaluate geometry results, we report Chamfer distance (CD), point-to-mesh distance (P2S), intersection over union (IOU), cosine normal consistency (NC-Cos), and L2 normal consistency (NC-L2). For all metrics, we sample every $100th$ frame from the test set, where the actor is performing unseen motions, and report the average. We provide an illustration of input cameras and evaluation cameras distribution in the supplementary material.

4.1   Comparisons

We compare our method with recent reconstruction and novel synthesis methods providing each of them wide-baseline four camera views as input during infer-

Table 1: Quantitative Comparison. Our method achieves state-of-the-art results for novel-view synthesis and geometry reconstruction. *Note, that ARAH requires 4D scans for meta learning and videos for the fine-tuning whereas other methods solely require static images.

Method	Subject	Appearance			Geometry				
		PSNR ↑	SSIM ↑	LPIPS ↓	NC-Cos ↓	NC-L2 ↓	Chamfer ↓	P2S ↓	IOU ↑
DeepMultiCap [94]	S2	-	-	-	0.143	0.445	1.305	1.303	0.764
	S27	-	-	-	0.149	0.454	1.466	1.831	0.797
	S5	-	-	-	0.183	0.504	2.238	2.224	0.794
DiffStereo [64]	S2	-	-	-	0.150	0.450	1.150	1.261	0.813
	S27	-	-	-	0.161	0.471	1.093	1.185	0.818
	S5	-	-	-	0.173	0.491	1.714	1.787	0.805
DVA [58]	S2	25.469	0.817	0.307	0.123	0.402	1.647	2.245	0.426
	S27	24.709	0.826	0.295	0.137	0.429	1.624	2.145	0.399
	S5	23.093	0.750	0.344	0.167	0.482	2.308	2.805	0.247
TransHuman [50]	S2	25.770	0.810	0.305	0.150	0.445	1.618	2.238	0.767
	S27	23.876	0.800	0.304	0.146	0.440	1.487	1.970	0.791
	S5	23.072	0.736	0.349	0.183	0.495	1.898	2.039	0.763
ARAH* [78]	S2	26.279	0.833	0.302	0.079	0.315	0.839	0.913	0.859
	S27	24.666	0.841	0.282	0.080	0.316	0.776	0.815	0.850
	S5	23.532	0.775	0.332	0.139	0.419	1.620	1.915	0.842
Ours	S2	26.529	0.841	0.249	0.096	0.351	0.679	0.814	0.887
	S27	25.284	0.849	0.247	0.104	0.370	0.614	0.734	0.891
	S5	23.996	0.777	0.302	0.147	0.448	1.133	1.277	0.877

Table 2: Quantitative Ablation. Here, we evaluate different initialization strategies and canonicalization types on rendering quality. Networks initialized with random or pre-trained weights tend to overfit the inputs, while our meta weights generalize well to novel views.

Init Type	Cano Type	Input View					Novel View				
		PSNR ↑	SSIM ↑	LPIPS ↓	NC-Cos ↓	NC-L2 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	NC-Cos ↓	NC-L2 ↓
Random	Root	31.678	0.955	0.113	0.173	0.455	23.802	0.782	0.276	0.171	0.454
	SMPL	31.180	0.953	0.120	0.225	0.535	23.691	0.770	0.292	0.214	0.523
	DDC	31.429	0.954	0.120	0.223	0.530	23.566	0.772	0.289	0.216	0.523
Pretrain	Root	-	-	-	-	-	-	-	-	-	-
	SMPL	-	-	-	-	-	-	-	-	-	-
	DDC	32.423	0.959	0.114	0.110	0.386	25.829	0.820	0.239	0.116	0.395
Meta	Root	24.477	0.870	0.253	0.170	0.480	23.094	0.764	0.326	0.168	0.475
	SMPL	28.224	0.903	0.215	0.092	0.345	26.120	0.833	0.258	0.096	0.349
	DDC	29.176	0.912	0.209	0.089	0.341	26.529	0.841	0.249	0.096	0.351

ence. Specially, we compare our method with generalizable reconstruction [64,94] and rendering [50] methods, a subject-specific rendering method [58], and a subject-specific optimization method with geometry prior initialization [78]. For more details, please refer to supplementary materials.

(1) Generalizable Methods. Given large-scale 3D human datasets [1, 94], DeepMultiCap [94] learns to predict occupancy fields from multi-view RGBs by fusing RGB features, normal features and voxel features with an attention layer, while DiffuStereo [64] trained a diffusion model to refine coarse stereo disparity. TransHuman [50] captures global relationships of human parts in the canonical space and learns to predict radiance fields from multi-view videos.

(2) Subject-specific Methods. Given dense-view RGB videos, DVA [58] learns articulated volumetric primitives attached to a parametric body model combined with texel-aligned features. ARAH [78] extends a meta-learned SDF prior [77]
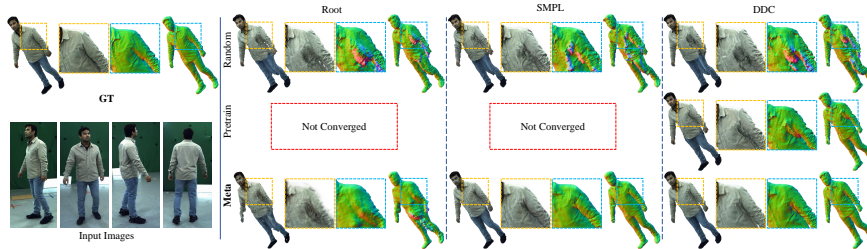
Fig. 5: Qualitative Ablation. We show results for different weight initialization strategies and space canonicalization types. Our meta-learning paradigm, together with the proposed space canonicalization, achieves the best result. We highlight that this applies irrespective of the choice of the character model (SMPL [39] or DDC [19]).

with a root-finding method to reconstruct clothed avatars from a sparse set of multi-view RGB videos.

We demonstrate qualitative and quantitative comparisons in Fig. 4 and Tab. 1.

Table 3: Quantitative Ablation. Here, we study initialization strategies and canonicalization types in terms of geometry results. Again, our meta-learning paradigm in combination with space canonicalization outperforms the baselines.

Init Type	Cano Type	Chamfer↓	P2S↓	IOU↑
	Root	1.610	2.253	0.823
Random	SMPL	1.291	1.596	0.806
	DDC	1.184	1.498	0.815
	Root	-	-	-
Pretrain	SMPL	-	-	-
	DDC	0.726	0.855	0.882
	Root	1.692	2.001	0.817
Meta	SMPL	0.799	0.897	0.885
	DDC	0.679	0.814	0.887

DeepMultiCap [94] directly regresses occupancy in world space, due to the articulated structure of humans and the limited scale of 3D scan datasets, their generalization ability is rather limited. DiffuStereo [64] struggles to find correct image correspondences under our challenging wide-baseline camera setup, leading to noisy reconstruction. DVA [58] learns rotation, translation, and scale of primitives to recover rough geometry. However, their texture un-projection step is sensitive to the template, and trac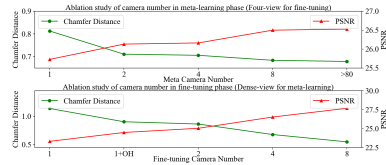king errors of the template lead to blurred results. TransHuman [50] integrates pixel-aligned features to improve texture details. However, they do not model geometry explicitly, which result in noisy geometry and texture boundaries. ARAH's geometry network [77] learns a meta prior taking 3D position as input and outputs the SDF value. However, due to their purely MLP-based architecture and non-end-to-end design, we found their results suffer from blurred appearance and less detailed geometry. Besides, the root-finding is not efficient taking over ten hours with 4 GPUs to finetune. In contrast, we use explicit hash encodings for spatial features, which is efficient and effective for meta feature learning. All the rendering methods above use parametric models [39, 53] as template, so they can not model loose clothing correctly (see S5 in Fig. 4). In contrast, our method recovers high-fidelity appearance and geometry even for loose types of apparel in minutes under this challenging setup, i.e. 4 wide-baseline cameras.

Fig. 6: Qualitative Ablation. A study of the influence of occlusion handling and the numbers of views during meta-learning (top) and fine-tuning (bottom).
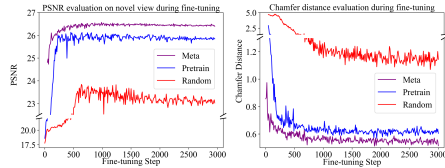

Fig. 7: Convergence and Quality. Note that our meta-learned weights achieve faster convergence and higher accuracy.

4.2 Ablation Studies

We perform ablation studies on the test set of the "S2" subject to evaluate the effectiveness of our design choices.

Weight Initialization and Space Canonicalization. Tab. 2/3 and Fig. 5 show the ablation results for weight initialization and space canonicalization. We choose the following baselines for weight initialization: *Random*: Network weights are initialized randomly. *Pretrain*: Network weights are pre-trained on all training frames. *Meta*: Network weights are learned through meta-learning. For space canonicalization, we choose the following baselines: *Root*: The space transformation is defined by the skeletal root rotation and translation. *SMPL*: The space transformation is defined by the nearest surface point on the SMPL [39] template. *DDC*: The space transformation is defined by the nearest point on the deformable DDC [19] template.

Random initialization typically overfits the input views, however, on novel views it performs poorly. Meta-learning in root-normalized space also performs poorly as the articulated nature of the humans poses a significant challenge for meta-learning the optimal network weights. In contrast, the combination of our key technical components, i.e. meta-learning weights *and* space canonicalization, effectively addresses those issues and achieves the best result for, both, geometry reconstruction and novel view synthesis. Moreover, we highlight that this holds true irrespective of the specific choice of the human template, i.e. SMPL or DDC.

Number of Camera Views and Occlusion Handling. Since our method solely takes videos for meta-learning and images for fine-tuning, it naturally supports arbitrary camera setups in *both* stages. As illustrated in Fig. 6, when increasing the number of cameras, our method obtains better rendering and reconstruction results. While this is not too surprising, we highlight that even under sparser settings our method performs reasonably well. Most interestingly, meta-learning can even be performed on a monocular video, which is not possible for ARAH [78] and MetaAvatar [77], since they rely on scans for meta-learning. Besides, at inference, our method allows fine-tuning using just a single view. Moreover, when performing monocular fine-tuning, our occlusion handling approach (1+OH) outperforms the baseline, which is not using this component.

Convergence Speed and Quality. In Fig. 7, we evaluate the convergence of our method compared to random and pre-trained weight initializations. Our

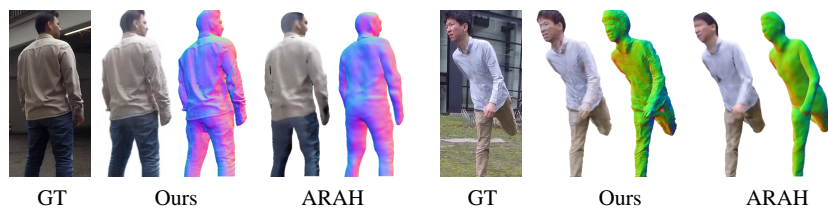GT   Ours   ARAH   GT   Ours   ARAH

Fig. 8: Qualitative In-the-wild Comparison. To evaluate the robustness to in-the-wild conditions, we compare our work and ARAH on WILDDYNACAP. Our approach achieves significantly better results in terms of appearance and geometric details.

proposed design outperforms the baselines in, both, convergence speed and accuracy. Since one fine-tuning step only takes $50ms$, our meta-learned initialization already converges after $40s$.

4.3   Evaluation on In-the-wild Sequences

Table 4: Quantitative In-the-wild Comparison. We outperform ARAH on in-the-wild sequences.

Method	Subject	PSNR ↑	SSIM ↑	LPIPS ↓
ARAH* [78]	S2	19.027	0.608	0.412
	S27	22.075	0.702	0.343
Ours	S2	19.156	0.661	0.351
	S27	22.080	0.750	0.260

Fig. 8 and Tab. 4 show results on our WILDDYNACAP dataset. Our method recovers high-fidelity geometry and appearance, consistently outperforming ARAH. These results show the robustness of our method to complex background, lighting conditions, and camera types.

5   Conclusion

This paper introduced METACAP, a novel approach towards high-fidelity performance capture and photo-realistic rendering of humans from very sparse multi-view or even monocular RGB images. We introduced a hybrid representation that benefits from both explicit and implicit human representation. Through comprehensive evaluations, we demonstrated that performing meta-learning with space canonicalization proves to be a crucial design factor, thereby providing a strong data-driven prior on the human capture. Our results demonstrate high-fidelity human reconstruction and free-view rendering as well as the versatility of approach in terms of camera setups, clothing types, and canonicalization strategies. In the future, we plan to explore even faster, potentially real-time, fine-tuning strategies as well as learning priors across individuals.

Limitations. Although our method achieves high-fidelity personalized human reconstruction and rendering from sparse observation, it still has a few limitations. First, our method can be sensitive to template fitting or motion capture results. Second, we do not take the temporal information into consideration. Integrating reliable observations or constraints from adjacent frames like Newcombe et al. [46] may lead to more robust geometry and rendering results. Third, our method can not handle hands very well. Modeling hand with a more fine-grained template and motion capture could be a potential solution.

Acknowledgements

References

1. https://web.twindom.com/
2. Antoniou, A., Edwards, H., Storkey, A.: How to train your maml. arXiv preprint arXiv:1810.09502 (2018)
3. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. ICCV (2021)
4. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. CVPR (2022)
5. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Zip-nerf: Anti-aliased grid-based neural radiance fields. ICCV (2023)
6. Bühler, M.C., Sarkar, K., Shah, T., Li, G., Wang, D., Helminger, L., Orts-Escolano, S., Lagun, D., Hilliges, O., Beeler, T., et al.: Preface: A data-driven volumetric prior for few-shot ultra high-resolution face synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3402–3413 (2023)
7. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: European Conference on Computer Vision (ECCV) (2022)
8. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14124–14133 (2021)
9. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. ACM Transactions on Graphics (ToG) **34**(4), 1–13 (2015)
10. Davydov, A., Remizova, A., Constantin, V., Honari, S., Salzmann, M., Fua, P.: Adversarial Parametric Pose Prior. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
11. De Luigi, L., Li, R., Guillard, B., Salzmann, M., Fua, P.: DrapeNet: Garment Generation and Self-Supervised Draping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
12. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International conference on machine learning. pp. 1126–1135. PMLR (2017)
13. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: CVPR (2022)
14. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. In: Proceedings of Machine Learning and Systems 2020, pp. 3569–3579 (2020)
15. Gu, J., Trevithick, A., Lin, K.E., Susskind, J., Theobalt, C., Liu, L., Ramamoorthi, R.: Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In: International Conference on Machine Learning (2023)
16. Guangcong, Chen, Z., Loy, C.C., Liu, Z.: Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. Technical Report (2023)

17. Guo, K., Lincoln, P., Davidson, P., Busch, J., Yu, X., Whalen, M., Harvey, G., Orts-Escolano, S., Pandey, R., Dourgarian, J., et al.: The relightables: Volumetric performance capture of humans with realistic relighting. ACM Transactions on Graphics (ToG) **38**(6), 1–19 (2019)
18. Habermann, M., Liu, L., Xu, W., Pons-Moll, G., Zollhoefer, M., Theobalt, C.: Hd-humans: A hybrid approach for high-fidelity digital humans. Proc. ACM Comput. Graph. Interact. Tech. (2023)
19. Habermann, M., Liu, L., Xu, W., Zollhoefer, M., Pons-Moll, G., Theobalt, C.: Real-time deep dynamic characters. ACM Transactions on Graphics **40**(4) (aug 2021)
20. Habermann, M., Xu, W., Zollhoefer, M., Pons-Moll, G., Theobalt, C.: Livecap: Real-time human performance capture from monocular video (2019)
21. Habermann, M., Xu, W., Zollhoefer, M., Pons-Moll, G., Theobalt, C.: Deepcap: Monocular human performance capture using weak supervision. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2020)
22. Hadwiger, M., Al-Awami, A.K., Beyer, J., Agus, M., Pfister, H.: Sparseleap: Efficient empty space skipping for large-scale volume rendering. IEEE transactions on visualization and computer graphics **24**(1), 974–983 (2017)
23. Hospedales, T., Antoniou, A., Micaelli, P., Storkey, A.: Meta-learning in neural networks: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(9), 5149–5169 (2021)
24. Huang, Y., Yi, H., Xiu, Y., Liao, T., Tang, J., Cai, D., Thies, J.: TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In: International Conference on 3D Vision (3DV) (2024)
25. Huber, P.J.: Robust estimation of a location parameter. In: Breakthroughs in statistics: Methodology and distribution, pp. 492–518. Springer (1992)
26. Jiang, W., Yi, K.M., Samei, G., Tuzel, O., Ranjan, A.: Neuman: Neural human radiance field from a single video. In: Proceedings of the European conference on computer vision (ECCV) (2022)
27. Jiang, Y., Habermann, M., Golyanik, V., Theobalt, C.: Hifecap: Monocular high-fidelity and expressive capture of human performances. In: BMVC (2022)
28. Johnson, E.C., Habermann, M., Shimada, S., Golyanik, V., Theobalt, C.: Unbiased 4d: Monocular 4d reconstruction with a neural deformation model. In: Computer Vision and Pattern Recognition Workshops (CVPRW) (2023)
29. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/
30. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
31. Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural human performer: Learning generalizable radiance fields for human performance rendering. Advances in Neural Information Processing Systems **34**, 24741–24752 (2021)
32. Kwon, Y., Liu, L., Fuchs, H., Habermann, M., Theobalt, C.: Deliffas: Deformable light fields for fast avatar synthesis. Advances in Neural Information Processing Systems (2023)
33. Li, K., Malik, J.: Learning to optimize. arXiv preprint arXiv:1606.01885 (2016)
34. Li, R., Tanke, J., Vo, M., Zollhöfer, M., Gall, J., Kanazawa, A., Lassner, C.: Tava: Template-free animatable volumetric actors. In: European Conference on Computer Vision. pp. 419–436. Springer (2022)
35. Li, Y., Habermann, M., Thomaszewski, B., Coros, S., Beeler, T., Theobalt, C.: Deep physics-aware inference of cloth deformation for monocular human perfor-

mance capture. In: 2021 International Conference on 3D Vision (3DV). pp. 373–384. IEEE Computer Society, Los Alamitos, CA, USA (dec 2021). `https://doi.org/10.1109/3DV53792.2021.00047`, `https://doi.ieeecomputersociety.org/10.1109/3DV53792.2021.00047`

36. Li, Z., Zheng, Z., Zhang, H., Ji, C., Liu, Y.: Avatarcap: Animatable avatar conditioned monocular human volumetric capture. In: European Conference on Computer Vision (ECCV) (October 2022)

37. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: Neural free-view synthesis of human actors with pose control. ACM Trans. Graph.(ACM SIGGRAPH Asia) (2021)

38. Long, X., Lin, C., Wang, P., Komura, T., Wang, W.: Sparseneus: Fast generalizable neural surface reconstruction from sparse views. ECCV (2022)

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

40. Luvizon, D., Golyanik, V., Kortylewski, A., Habermann, M., Theobalt, C.: Relightable neural actor with intrinsic decomposition and pose control. In: European Conference on Computer Vision (ECCV) (2024)

41. Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., Black, M.J.: Learning to Dress 3D People in Generative Clothing. In: Computer Vision and Pattern Recognition (CVPR) (2020)

42. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4460–4470 (2019)

43. Mihajlovic, M., Bansal, A., Zollhoefer, M., Tang, S., Saito, S.: KeypointNeRF: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In: European conference on computer vision (2022)

44. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)

45. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)

46. Newcombe, R.A., Fox, D., Seitz, S.M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 343–352 (2015)

47. Nichol, A., Achiam, J., Schulman, J.: On first-order meta-learning algorithms. arXiv preprint arXiv:1803.02999 (2018)

48. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5480–5490 (2022)

49. Palafox, P., Sarafianos, N., Tung, T., Dai, A.: Spams: Structured implicit parametric models. CVPR (2022)

50. Pan, X., Yang, Z., Ma, J., Zhou, C., Yang, Y.: Transhuman: A transformer-based human representation for generalizable neural human rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3544–3555 (October 2023)

51. Pang, H., Zhu, H., Kortylewski, A., Theobalt, C., Habermann, M.: Ash: Animatable gaussian splats for efficient and photoreal human rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1165–1175 (June 2024)
52. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 165–174 (2019)
53. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019)
54. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: ICCV (2021)
55. Peng, S., Zhang, S., Xu, Z., Geng, C., Jiang, B., Bao, H., Zhou, X.: Animatable neural implicit surfaces for creating avatars from videos. arXiv preprint arXiv:2203.08133 (2022)
56. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: CVPR (2021)
57. Rajeswaran, A., Finn, C., Kakade, S.M., Levine, S.: Meta-learning with implicit gradients. Advances in neural information processing systems **32** (2019)
58. Remelli, E., Bagautdinov, T., Saito, S., Wu, C., Simon, T., Wei, S.E., Guo, K., Cao, Z., Prada, F., Saragih, J., et al.: Drivable volumetric avatars using texel-aligned features. In: ACM SIGGRAPH 2022 Conference Proceedings (2022)
59. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2304–2314 (2019)
60. Saito, S., Simon, T., Saragih, J., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 84–93 (2020)
61. Saito, S., Yang, J., Ma, Q., Black, M.J.: SCANimate: Weakly supervised learning of skinned clothed avatar networks. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (Jun 2021)
62. Shao, R., Chen, L., Zheng, Z., Zhang, H., Zhang, Y., Huang, H., Guo, Y., Liu, Y.: Floren: Real-time high-quality human performance rendering via appearance flow using sparse rgb cameras. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–10 (2022)
63. Shao, R., Zhang, H., Zhang, H., Chen, M., Cao, Y., Yu, T., Liu, Y.: Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In: CVPR (2022)
64. Shao, R., Zheng, Z., Zhang, H., Sun, J., Liu, Y.: Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In: European Conference on Computer Vision. pp. 702–720. Springer (2022)
65. Shen, K., Guo, C., Kaufmann, M., Zarate, J., Valentin, J., Song, J., Hilliges, O.: X-avatar: Expressive human avatars. Computer Vision and Pattern Recognition (CVPR) (2023)

66. Shetty, A., Habermann, M., Sun, G., Luvizon, D., Golyanik, V., Theobalt, C.: Holoported characters: Real-time free-viewpoint rendering of humans from sparse rgb cameras (2023)
67. Shuai, Q., Geng, C., Fang, Q., Peng, S., Shen, W., Zhou, X., Bao, H.: Novel view synthesis of human interactions from sparse multi-view videos. In: SIGGRAPH Conference Proceedings (2022)
68. Sitzmann, V., Chan, E., Tucker, R., Snavely, N., Wetzstein, G.: Metasdf: Meta-learning signed distance functions. Advances in Neural Information Processing Systems **33**, 10136–10147 (2020)
69. Stoll, C., Hasler, N., Gall, J., Seidel, H.P., Theobalt, C.: Fast articulated motion tracking using a sums of gaussians body model. In: 2011 International Conference on Computer Vision. pp. 951–958. IEEE (2011)
70. Su, Z., Xu, L., Zheng, Z., Yu, T., Liu, Y., Fang, L.: Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. pp. 246–264. Springer (2020)
71. Sun, G., Chen, X., Chen, Y., Pang, A., Lin, P., Jiang, Y., Xu, L., Yu, J., Wang, J.: Neural free-viewpoint performance rendering under complex human-object interactions. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 4651–4660 (2021)
72. Tancik, M., Mildenhall, B., Wang, T., Schmidt, D., Srinivasan, P.P., Barron, J.T., Ng, R.: Learned initializations for optimizing coordinate-based neural representations. In: CVPR (2021)
73. Tretschk, E., Kairanda, N., B R, M., Dabral, R., Kortylewski, A., Egger, B., Habermann, M., Fua, P., Theobalt, C., Golyanik, V.: State of the art in dense monocular non-rigid 3d reconstruction. Computer Graphics Forum (Eurographics State of the Art Reports) (2023)
74. Wang, K., Peng, S., Zhou, X., Yang, J., Zhang, G.: Nerfcap: Human performance capture with dynamic neural radiance fields. IEEE Transactions on Visualization and Computer Graphics (2022)
75. Wang, K., Zhang, G., Cong, S., Yang, J.: Clothed human performance capture with a double-layer neural radiance fields. In: Computer Vision and Pattern Recognition (CVPR) (2023)
76. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. NeurIPS (2021)
77. Wang, S., Mihajlovic, M., Ma, Q., Geiger, A., Tang, S.: Metaavatar: Learning animatable clothed human models from few depth images. In: Advances in Neural Information Processing Systems (2021)
78. Wang, S., Schwarz, K., Geiger, A., Tang, S.: Arah: Animatable volume rendering of articulated human sdfs. In: European Conference on Computer Vision (2022)
79. Wang, Y., Han, Q., Habermann, M., Daniilidis, K., Theobalt, C., Liu, L.: Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. arXiv preprint arXiv:2212.05231 (2022)
80. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16210–16220 (June 2022)
81. Xiang, D., Prada, F., Cao, Z., Guo, K., Wu, C., Hodgins, J., Bagautdinov, T.: Drivable avatar clothing: Faithful full-body telepresence with dynamic clothing

driven by sparse rgb-d input. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–11 (2023)

82. Xiang, D., Prada, F., Wu, C., Hodgins, J.: Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In: 2020 International Conference on 3D Vision (3DV). pp. 322–332. IEEE (2020)

83. Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J.: Econ: Explicit clothed humans optimized via normal integration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 512–523 (2023)

84. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: ICON: Implicit Clothed humans Obtained from Normals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13296–13306 (June 2022)

85. Xu, W., Chatterjee, A., Zollhöfer, M., Rhodin, H., Mehta, D., Seidel, H.P., Theobalt, C.: Monoperfcap: Human performance capture from monocular video. ACM Trans. Graph. **37**(2), 27:1–27:15 (May 2018). `https://doi.org/10.1145/3181973`, `http://doi.acm.org/10.1145/3181973`

86. Xue, Y., Bhatnagar, B.L., Marin, R., Sarafianos, N., Xu, Y., Pons-Moll, G., Tung, T.: NSF: Neural Surface Field for Human Modeling from Monocular Depth. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2023)

87. Yang, J., Pavone, M., Wang, Y.: Freenerf: Improving few-shot neural rendering with free frequency regularization (2023)

88. Yu, A., Li, R., Tancik, M., Li, H., Ng, R., Kanazawa, A.: PlenOctrees for real-time rendering of neural radiance fields. In: ICCV (2021)

89. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelNeRF: Neural radiance fields from one or few images. In: CVPR (2021)

90. Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., Liu, Y.: Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7287–7296 (2018)

91. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595. IEEE Computer Society, Los Alamitos, CA, USA (2018). `https://doi.org/10.1109/CVPR.2018.00068`, `https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00068`

92. Zhao, F., Jiang, Y., Yao, K., Zhang, J., Wang, L., Dai, H., Zhong, Y., Zhang, Y., Wu, M., Xu, L., et al.: Human performance modeling and rendering via neural animated mesh. ACM Transactions on Graphics (TOG) **41**(6), 1–17 (2022)

93. Zhao, F., Yang, W., Zhang, J., Lin, P., Zhang, Y., Yu, J., Xu, L.: Humannerf: Efficiently generated human radiance field from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7743–7753 (June 2022)

94. Zheng, Y., Shao, R., Zhang, Y., Yu, T., Zheng, Z., Dai, Q., Liu, Y.: Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6239–6249 (2021)

95. Zheng, Z., Huang, H., Yu, T., Zhang, H., Guo, Y., Liu, Y.: Structured local radiance fields for human avatar modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2022)

96. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction (2021)

97. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)

98. Zhu, H., Zhan, F., Theobalt, C., Habermann, M.: Trihuman : A real-time and controllable tri-plane representation for detailed human geometry and appearance synthesis (2023)

99. Zuo, X., Wang, S., Zheng, J., Yu, W., Gong, M., Yang, R., Cheng, L.: Sparsefusion: Dynamic human avatar modeling from sparse rgbd images. IEEE Transactions on Multimedia **23**, 1617–1629 (2020)