DOI: 10.1111/cgf.14774

EUROGRAPHICS 2023
A. Bousseau and C. Theobalt
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 42 (2023), Number 2
STAR – State of The Art Report

State of the Art in Dense Monocular Non-Rigid 3D Reconstruction

Edith Tretschk[1⋆] Navami Kairanda[1⋆] Mallikarjun B R[1] Rishabh Dabral[1] Adam Kortylewski[1,2]
Bernhard Egger[3] Marc Habermann[1] Pascal Fua[4] Christian Theobalt[1] Vladislav Golyanik[1]

[1]Max Planck Institute for Informatics, Saarland Informatics Campus [2]University of Freiburg
[3]Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) [4]EPFL ⋆denotes equal contribution

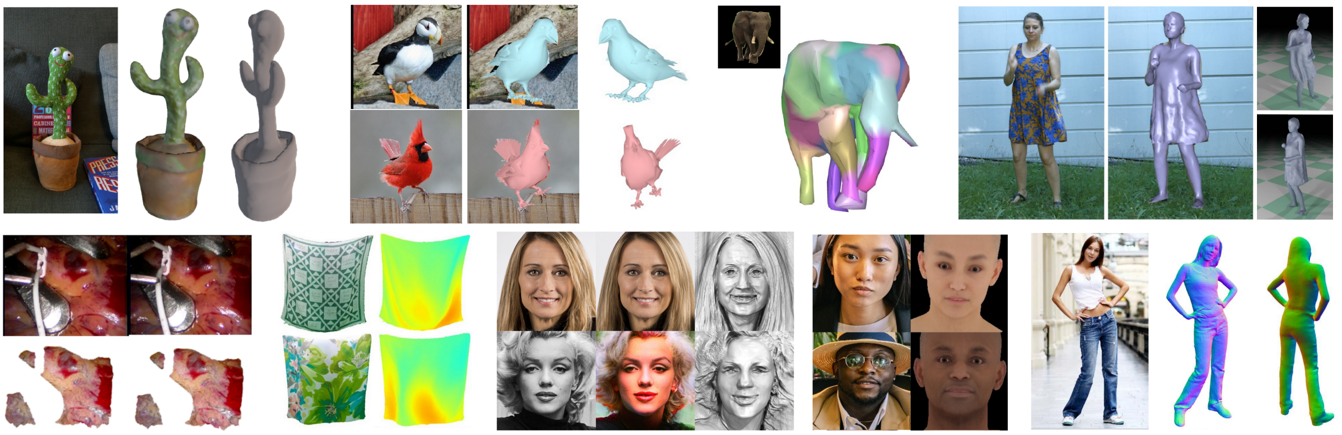

Figure 1: *We review state-of-the-art methods for the dense 3D reconstruction of deformable objects from monocular images and videos, such as general deformable surfaces, soft body tissues in medical scenarios, animals, and human bodies and body parts. Images adapted from [JHS*22, WKDB21, YSJ*21b, HXZ*20, STG*20, KTE*22, CLC*22, FBT*22, SSSJ20].*

Abstract

3D reconstruction of deformable (or non-rigid*) scenes from a set of monocular 2D image observations is a long-standing and actively researched area of computer vision and graphics. It is an ill-posed inverse problem, since—without additional prior assumptions—it permits infinitely many solutions leading to accurate projection to the input 2D images. Non-rigid reconstruction is a foundational building block for downstream applications like robotics, AR/VR, or visual content creation. The key advantage of using monocular cameras is their omnipresence and availability to the end users as well as their ease of use compared to more sophisticated camera set-ups such as stereo or multi-view systems. This survey focuses on state-of-the-art methods for dense non-rigid 3D reconstruction of various deformable objects and composite scenes from monocular videos or sets of monocular views. It reviews the fundamentals of 3D reconstruction and deformation modeling from 2D image observations. We then start from general methods—that handle arbitrary scenes and make only a few prior assumptions—and proceed towards techniques making stronger assumptions about the observed objects and types of deformations (e.g. human faces, bodies, hands, and animals). A significant part of this STAR is also devoted to classification and a high-level comparison of the methods, as well as an overview of the datasets for training and evaluation of the discussed techniques. We conclude by discussing open challenges in the field and the social aspects associated with the usage of the reviewed methods.*

1. Introduction

Humans can close one eye, look around, and get a fair sense of their surroundings in terms of their 3D geometry, appearance, and even deformations [Sch86, WJ97]. Nevertheless, designing computational methods that densely reconstruct a dynamic scene in 3D using a single monocular camera remains a challenging task that is far from solved, as this STAR shows.

Monocular 3D reconstruction is a challenging domain of computer vision and graphics motivated by fundamental questions and practical applications. The rigid case has been studied for decades and mature methods are available nowadays [LH87, AFS*11, MAMT15]; the rigidity assumption, *i.e.,* that the transformation can be entirely described by a single translation and a single rotation (6DoF), significantly simplifies the formulation compared to the non-rigid case (>6DoF, often ≫6DoF) and provides strong prior knowledge about the expected 3D structure. At the same time, while some objects preserve their states longer than others, all eventually deform over time [Lar22] while being exposed to physical forces. We thus live in a constantly changing world, irrespective of the scale: The scale of our galaxy, the solar system, Earth, ecosystems on our planet, individual living species, humans, human body parts (*e.g.* face and hands) and organs (*e.g.* heart), cells or atoms. Many spectacular dynamic effects are inherently non-rigid.

There is much interest in monocular approaches both in the computer vision and graphics communities, as evidenced by the many published works in a wide range of domains, *e.g.* monocular depth estimation [GBCR16], image segmentation [LXWY22], or image synthesis [KLA19]. Since RGB cameras are ubiquitous and single-camera setups are much easier to deploy than multi-camera ones, monocular methods are relevant not just out of sheer curiosity about the limits of reconstruction under the most challenging conditions but also because they enable a multitude of applications. Applications for monocular 3D reconstruction range from geometry acquisition [AFS*11], novel view synthesis [TTG*21] and elastic parameter estimation [KTE*22] to scene or video editing [GZC*16] and scene recognition and understanding [COR*16]. All these applications are highly relevant for such fields of science and engineering as VR/AR, movie and game production, content creation, computer-assisted design, cultural heritage, robotics, space exploration, experimental physics, medicine, zoology and many others. In other words, reliable solutions to monocular 3D reconstruction have the potential to impact society in significant ways.

This is the first STAR devoted to non-rigid 3D reconstruction from single monocular cameras (we discuss related surveys in Sec. 1.2). In recent years, monocular 3D reconstruction has been reinvigorated by several breakthroughs, including widely applicable parametric models [PCG*19, QWM*20, LL21]; neural parametrizations [MST*20]; machine learning techniques [KTEM18]; high-quality, large-scale datasets [WBW*11, LLWT15, MYW*20]; and powerful computational resources, to name a few. Thus, monocular reconstruction methods nowadays produce 3D outputs of impressive visual quality that are suitable for many applications discussed above, including computer graphics; see Fig. 1 for representative 3D reconstructions by state-of-the-art methods. Even a few years back, this was not generally true. Still, despite great progress, there remain a lot of unsolved problems in monocular non-rigid 3D reconstruction. Since the field has recently undergone massive change, we seek to document its current state and the challenges researchers will face during the upcoming years.

1.1. Scope of this STAR

This STAR focuses on methods from recent years for non-rigid 3D reconstruction that take one or several consecutive views from a single camera as input and that output dense 3D reconstructions of the scene in each view or point in time spanning the observations. We put a special emphasis on the emerging fields of neural scene representations and neural rendering, physics-based reconstruction, and reconstruction from event cameras. We next explain the meaning of each core word of this STAR's title in more detail.

Dense. We focus on dense 3D reconstructions and leave the sparse case out of scope for several reasons: 1) Dense reconstructions provide a more complete scene description; 2) Nowadays, sufficient computational resources and the availability of reliable dense preprocessing methods allow many downstream applications to assume dense deformable 3D scenes; 3) Many principles are shared between the dense and sparse cases and, hence, most fundamentals we discuss in Sec. 2 cover both; and 4) Considering the literature volume in the field, even a survey of a format like this one cannot cover both cases with satisfactory depth.

Monocular. We only consider methods where no more than a single view observes each 3D scene state (no multi-view). We mainly focus on sensors that register incoming light in the visible spectrum (wavelengths in the range 320−1100 nm). Thus, event cameras are in scope of this STAR (see Sec. 2.4) but active sensor systems with active emitters such as RGB-D cameras are not. However, we only apply these criteria at test time, and hence any supervision (including 3D) at training time is inside the scope.

Non-Rigid. We only consider objects that can deform. We cover methods for static (3D; single timestep) and dynamic (4D; multiple timesteps) reconstruction. Many approaches parametrize non-rigid deformations by statistical 3D models. Widely-used parametric human body models neglect different clothing styles, facial expressions and hairstyles, and only provide shape proxies, *i.e.* approximate shapes that do not allow recognizing a person from the reconstructed geometry. We thus believe it is not enough to instantiate a parametric human body model to claim that an approach reconstructs dense 3D human shapes. Hence, methods that do not perform geometric refinement and add identity-specific characteristics on top of shape proxies are out of our scope. The situation is different with human hands, human faces, and animals. Hands are mostly observed naked, are easier to capture, and vary less across people. Similar statements apply to faces. Next, there is little work on dense 3D animal reconstruction from monocular views. Hence, we include methods that use parametric hand, face, and animal models.

3D. We focus on true 3D representations and ignore image-based (*e.g.* 2.5D/depth) or intermediate representations (*e.g.* light fields).

Reconstruction. We seek a model that ideally represents the scene as it was observed. We do not require it to be generative or editable.

1.2. Related Surveys

Several method surveys and STARs were published over the last twelve years; some of them are outdated as of 2022. Salzmann and Fua [SF10] review methods for deformable 3D surface reconstruction, only a few of which addressed the dense case in 2010. Jensen *et al.* [HBAD21] review Non-Rigid Structure-from-Motion (NRSfM) and introduce the sparse *NRSfM 2017 challenge dataset*. They focus on sparse techniques from before 2020; dense NRSfM techniques are not systematically discussed unless evaluated on the proposed dataset. A recent short survey [KPL*22] reviews generalizable deep-learning methods for dense 3D reconstruction of rigid and non-rigid objects from a single image, with weak prior knowledge about the object class. Our report is much more exhaustive, including category-specific methods and volume-rendering techniques. A STAR by Zollhoefer *et al.* [ZSG*18] focuses on rigid and non-rigid 3D reconstruction from RGB-D cameras, which are out of our scope. STARs on neural rendering [TFT*20, XTS*22, TTM*22] focus on novel view synthesis of rigid and non-rigid scenes. They cover a small subset of techniques that we cover. Other surveys are devoted to only face [ZTG*18, EST*20] or bird reconstruction [MJK*22]. Tian *et al.* [TZLW22] cover monocular 3D human mesh recovery using parametric models. In contrast, we focus on monocular methods that can regress human shapes beyond shape proxies and naked humans. An unpublished survey [XX22] discusses 3D-aware image synthesis methods but covers only a few of the non-rigid methods covered by our survey. A recent survey [GDO*22] on event-based vision covers *static* simultaneous localization and mapping (SLAM) from event cameras, while we discuss *non-rigid* event-based methods.

All in all, the STAR at hand is the first one that systematically reviews all types of monocular dense non-rigid 3D reconstruction techniques for various scenes and objects (together with the fundamentals for introducing the field or catching up with the field), whereas previous surveys address only small parts of this report.

1.3. Paper Selection Criteria

We predominantly discuss works from international computer vision and graphics conferences and journals that are in scope of this STAR (*cf.* Sec. 1.1). We also include a few recent technical reports on arXiv.org. However, considering how fast the field is developing, we cannot claim completeness in either case.

1.4. Structure of this STAR

We first motivate this STAR in this introductory Sec. 1. We next describe the basics of non-rigid 3D reconstruction in Sec. 2, which covers many aspects that are useful in order to understand any particular work on non-rigid 3D reconstruction. At the core of our report is the discussion of the current state of the art in Sec. 3, which is ordered by the object category that is to be reconstructed. We discuss cross-sectional aspects and open challenges in Sec. 4. Finally, we provide an overview of the impact that this field has on society in Sec. 5, and draw conclusions in Sec. 6.

2. Fundamentals

This section describes the main building blocks of the design pipeline of non-rigid 3D reconstruction methods. We aim to provide a guide to the reader of all the pieces involved, making a critical reading of recent works possible. However, we do not claim full coverage of all aspects. For example, we focus on the basics of computer graphics and computer vision required to understand this STAR, and we assume pre-requisite knowledge in machine learning on the part of the reader. We first take a functional look at the components of 3D reconstruction in Sec. 2.1, which we then use in Sec. 2.2 to formulate the reconstruction problem that we are concerned with. In Sec. 2.3, we describe how to parametrize the functions discussed earlier. Sec. 2.5 describes data terms that are commonly used to obtain consistency between the input and the model parametrization. Sec. 2.6 discusses multiple challenges that we face when trying to obtain a solution. Then, in Sec. 2.7, we specifically look at the underconstrained nature of the problem and provide a high-level description of several standard priors to tackle it. Finally, Sec. 2.8 shortly describes the optimization of the solution parameters with respect to the resulting loss function.

2.1. Background: A Functional Look

In this section, we introduce basic concepts from computer graphics [FVVD*96]: geometry, deformations, appearance, and rendering. For didactic purposes, we explicitly split the commonly used term *representation* into its two constitutive concepts of *function* and *parametrization*. While this section takes a theoretical, functional perspective, we present practical parametrizations in a later section.

Notation. We use $\mathbf{x} \in \mathbb{R}^n$ for a vector, $\mathbf{A} \in \mathbb{R}^{m \times n}$ for a matrix, S for a set, $f : S \to T$ for a function from S to T, $\times$ for the cross product. We use d for deformations; subscripts, *e.g.* d_t, to denote time t; vertex index i in triangulations. Time derivatives have dots on top: $\dot{\mathbf{x}}, \ddot{\mathbf{x}}$.

2.1.1. Geometry Functions

We first require a representation of the 3D geometry of the object or scene. The most common way of representing an object's geometry is via its 2D surface $S \subset \mathbb{R}^3$. A 2D surface can be described *implicitly*, *i.e.* defined on a volumetric/3D domain, via an *indicator function* $s : \mathbb{R}^3 \to \{0, 1\}$ that is 1 on the surface and 0 otherwise:

$$s(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in S \\ 0 & \text{else} \end{cases}. \tag{1}$$

A level-set function provides additional information about where the surface can be found:

$$s(\mathbf{x}) = \min_{\mathbf{y} \in S} \|\mathbf{x} - \mathbf{y}\|_2. \tag{2}$$

This *unsigned distance function (UDF)* specifies, for each point in 3D space, how far away it is from the closest surface. The surface lies at the 0-level set: $\{\mathbf{x} | s(\mathbf{x}) = 0\}$. This can be turned into the *signed distance function (SDF)* by giving points inside the object negative distance, *i.e.* for $\mathbf{x}$ inside the object, $SDF(\mathbf{x}) = -UDF(\mathbf{x})$. Thus, an SDF is only defined for closed surfaces.

In contrast to these implicit surface representations, we can make use of the mathematical definition of a 2D manifold embedded in 3D space. A manifold admits an *explicit* surface description: a map from a subset U of 2D space to 3D space (by embedding all charts of the surface atlas into the same $\mathbb{R}^2$ space). The resulting surface of such a UV map, $UV : U \subset \mathbb{R}^2 \to \mathbb{R}^3$, is called a *parametric surface*:

$$s(\mathbf{u}) = UV(\mathbf{u}). \tag{3}$$

We sometimes want to model objects without forcing a clearly defined surface on them (*e.g.* smoke) or without putting the surface at the center of the parametrization (*e.g.* fluids). In such cases, volumetric representations encode geometry in a soft manner. Density fields $v : \mathbb{R}^3 \to [0, +\infty)$ are the most common volumetric function, where a density of 0 denotes empty space:

$$v(\mathbf{x}) = \text{density}(\mathbf{x}). \tag{4}$$

© 2023 The Authors.
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

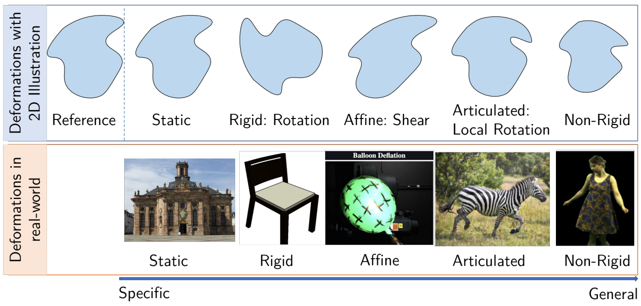

Figure 2: *Different types of deformations, in order of expressiveness. Top: Illustration of the deformation map from the reference to the deformed geometry. Bottom: Example objects commonly studied in computer vision. Images adapted from [RES*22, CFG*15, HBAD21, YHL*22, HXZ*20].*

Furthermore, applying volumetric representations to surfaces allows for more slack in the optimization [MST*20].

2.1.2. Deformation Functions

We next specify different types of deformations that can be applied to a geometry, with each one generalizing the previous type, see Fig. 2. We define the deformed geometry g_d with respect to an undeformed template or reference geometry g, where *e.g.* $g = s$ for a surface.

Static objects do not move locally or globally, with the deformed geometry g_d trivially defined as $g_d(\mathbf{x}) = g(\mathbf{x})$.

Rigid objects may move around globally, without changing their shape and size. Mathematically, the deformed geometry g_d can rotate and translate: $g_d(\mathbf{x}) = g(\mathbf{R}\mathbf{x}+\mathbf{t})$, for a 3D rotation matrix $\mathbf{R} \in SO(3)$ and a 3D translation $\mathbf{t} \in \mathbb{R}^3$. This is an idealization where the deformation is so small that it can be neglected.

Affine deformations are also global like rigid deformations, except that now shearing and scaling are allowed: $g_d(\mathbf{x}) = g(\mathbf{A}\mathbf{x}+\mathbf{t})$, where $\mathbf{A} \in \mathbb{R}^{3\times 3}$ is an invertible linear map.

Articulated deformations generalize the previous classes to (piecewise) ensembles of N local rigid/affine deformations $\{d_i\}_{i=1}^N$, where $d_i(\mathbf{x}) = \mathbf{R}_i\mathbf{x} + \mathbf{t}_i$ in the rigid case:

$$g_d(\mathbf{x}) = g(d_i(\mathbf{x})) \quad \text{if } d_i(\mathbf{x}) \in U_i, \tag{5}$$

where $\{U_i \subset \mathbb{R}^3\}_{i=1}^N$ is a partition of $\mathbb{R}^3$. U_i is a *local part* and deforms according to its own associated deformation d_i. Humans or animals are sometimes modeled with articulated deformations.

Non-Rigid objects undergo elastic deformations as they respond naturally to applied forces, constraints and contacts with self or obstacles. This most generic formulation describes the behavior of most real, physical objects and deformation d is any (in general non-linear) map that displaces an undeformed point to a deformed one:

$$g_d(\mathbf{x}) = g(d(\mathbf{x})), \text{ where } d : \mathbb{R}^3 \to \mathbb{R}^3. \tag{6}$$

Deformation Measures can be used to quantify the geometric amount of deformation and they often derive from differential geometry of solids, surfaces, and curves [DC16]. As they measure deformations, they need to be invariant under rigid transformations. We later derive deformation constraints from them that act as priors, see Sec. 2.7.

Let $\mathbf{m}$ be the *intrinsic* or *material coordinates* of a point in an undeformed object. For volumes, surfaces, and curves, we have $\mathbf{m} \in \mathbb{R}^3$, $\mathbf{m} \in \mathbb{R}^2$, and $\mathbf{m} \in \mathbb{R}$, respectively. We can map the point from its material position to the world-space position $\mathbf{x} = \mathbf{x}(\mathbf{m}) \in \mathbb{R}^3$. The shape of any such object is determined by the Euclidean distances between nearby world-space points. Non-rigid deformations may change these distances, while rigid deformations (and reflections) do not. If $\mathbf{m}$ and $\mathbf{m} + d\mathbf{m}$ are the material coordinates of two nearby points, the differential length dl between them in the deformed object is:

$$dl = \sum_{i,j} G_{ij} dm_i dm_j, \tag{7}$$

where the symmetric and positive definite matrix $G \in \mathbb{R}^{3\times 3}$, *i.e.*

$$G_{ij}(\mathbf{x}(\mathbf{m})) = \frac{\partial \mathbf{x}}{\partial m_i} \cdot \frac{\partial \mathbf{x}}{\partial m_j} \tag{8}$$

is known as the *metric tensor* or the *first fundamental form*. Two volumetric objects have the same shape (up to a rigid motion) and considered *isometric* if their metric tensors are identical functions of $\mathbf{m}$ everywhere. Isometry is not a sufficient condition for rigidity of surfaces and curves which are volumes infinitesimally thin in one or two dimensions. Surfaces can change shape by bending, even while preserving geodesic distances between nearby points. While the metric tensor describes the in-plane stretching and shearing of surfaces, the *curvature tensor* $B \in \mathbb{R}^{2\times 2}$, or *second fundamental form*, quantifies bending and is defined as:

$$B_{ij}(\mathbf{x}(\mathbf{m})) = \mathbf{n} \cdot \frac{\partial^2 \mathbf{x}}{\partial m_i \partial m_j}, \tag{9}$$

where $\mathbf{n}$ is the unit surface normal. Together, they form the *shape operator* $G^{-1}B$. Two surfaces have the same shape if their first and second fundamental forms are identical.

Deformation Dynamics of an object occur under applied forces. The dynamics are determined by the object's initial shape and its material configuration. Continuum mechanics [SB12] and elasticity theory [Sla12] formulate quantitative descriptions for the deformation of a continuous object. One can then arrive at partial differential equations (PDEs) that model the dynamic behaviors of objects.

Let the time-varying *deformation map* be given as $\mathbf{x} = \mathbf{x}(\mathbf{m}, t)$, with $\mathbf{X} \equiv \mathbf{x}(\mathbf{m}, 0)$ as the undeformed *reference configuration*. An important physical quantity derived from $\mathbf{x}$ is the *deformation gradient* $\mathbf{F} \equiv \nabla_{\mathbf{m}}\mathbf{x} \in \mathbb{R}^{3\times 3}$, the spatial Jacobian of the deformation map. The metric tensor $G \equiv \mathbf{F}^T\mathbf{F}$ provides a measure for local distortion of lengths and angles relative to the reference shape (and connects the physical response described here with the geometric deformations described in the previous subsection). We can derive strain measures from the deformation gradient to quantify the geometric severity of the deformation. Assuming an identity metric tensor for the reference shape, the *Green strain tensor* $\varepsilon \in \mathbb{R}^{3\times 3}$, a

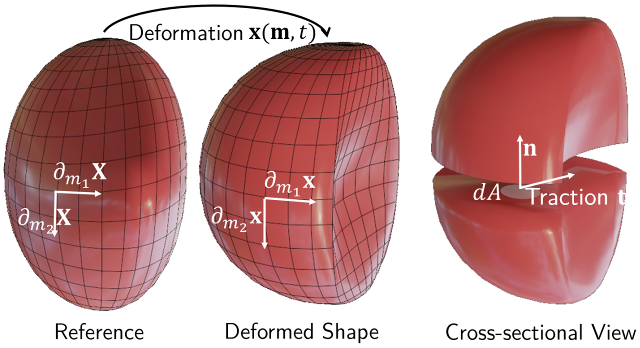
Deformation $\mathbf{x}(\mathbf{m},t)$
$\partial_{m_1}\mathbf{X}$
$\partial_{m_2}\mathbf{X}$
$\partial_{m_1}\mathbf{x}$
$\partial_{m_2}\mathbf{x}$
$\mathbf{n}$
dA
Traction $\mathbf{t}$
Reference
Deformed Shape
Cross-sectional View

Figure 3: *Local tangents on the volume can be used to measure the deformation between the reference* $\mathbf{X}$ *and the deformed state* $\mathbf{x}$*. The traction* $\mathbf{t}$ *is the density of the internal forces on the area dA with normal* $\mathbf{n}$*.*

commonly used strain measure, is given as:

$$\varepsilon = \frac{1}{2}(\mathbf{F}^T\mathbf{F} - \mathbf{I}). \tag{10}$$

This strain omits information unrelated to shape change from $\mathbf{F}$ but retains information about the local deformation magnitude.

As a result of elastic deformation, the objects accumulate potential energy and the resulting internal elastic forces are often described by the *Cauchy stress tensor* $\sigma \in \mathbb{R}^{3\times3}$. Constitutive models relate the (geometric) strain to the (physical) material response it triggers, such as force, stress, or strain energy. In its most general form, the constitutive equation is formulated as *Hooke's law*:

$$\sigma_{ij} = \mathbf{C}_{ijkl}\varepsilon_{kl}, \tag{11}$$

where $\mathbf{C}$ is a rank-4, possibly non-linear elastic tensor.

The distribution of internal elastic forces that result from a deformation can be described as follows: Consider a slice of the deformable object with differential surface area dA and normal $\mathbf{n}$; see Fig. 3. Then the traction $\mathbf{t}$ along the cut is the (surface) force density function that measures the force per unit undeformed area:

$$\mathbf{t} = \lim_{dA\to0}\frac{d\mathbf{f}}{dA} = \sigma\mathbf{n}, \tag{12}$$

where the Cauchy stress tensor σ serves as a fundamental force descriptor that generalizes traction for every normal direction $\mathbf{n}$ and relates the internal forces to the deformations using (11).

Consider now a volume element V of a deformable object with boundary surface ∂V. Let $\mathbf{f}$ be the external forces such as gravity, wind per unit volume acting on V. Balance of linear momentum postulates that the resultant of the external forces acting on the object is equal to the rate of change of its total linear momentum:

$$\int_{\partial V}\mathbf{t}\,da + \int_V \mathbf{f}\,dv = \int_V \rho\ddot{\mathbf{x}}\,dv, \tag{13}$$

where ρ is the mass density and $\ddot{\mathbf{x}}$ is the material acceleration, which together constitute the inertial forces. The divergence theorem lets us change the surface integral in (13) into a volume integral:

$$\int_{\partial V}\mathbf{t}\,da = \int_{\partial V}\sigma\mathbf{n}\,da = \int_V \text{div}\,\sigma\,dv. \tag{14}$$

As (13) must hold for any enclosed volume, the point-wise equation of motion, the so-called strong form, follows as:

$$\text{div}\,\sigma(\mathbf{x},\dot{\mathbf{x}}) + \mathbf{f} = \rho\ddot{\mathbf{x}}, \tag{15}$$

after accounting for velocity-dependent damping forces.

This 3D case gives rise to specialized theories when one of the dimensions becomes very small; for example, the continuum mechanics of 2D surfaces are given by the classical *Kirchhoff-Love shell theory* [WTP03], while two very small dimensions lead to beam theory, for example *Euler-Bernoulli beam theory* [ZTTT00] for 1D curves.

2.1.3. Appearance Functions

Apart from geometries along with their deformations, a 3D scene description in computer graphics requires specifying the lights and the material models. While geometry captures the macro-structure of an object or scene, material is determined by the object's micro-structure. Then, the interaction of the geometry with the lights determines its (surface) appearance towards a camera. Consequently, physically-based simulation of light transport forms the basis for rendering. Based on the material composition and surface roughness, same geometry can reflect light differently and thus have different surface appearance [LGC*05]. The commonly used *Lambertian* or *diffuse* material refers to rough surfaces, where light is reflected multiple times within the material that it loses directionality. It thus does not vary with the viewing angle. A smoother surface generates *glossy* appearance while perfectly smooth one leads to a *specular (mirror)* reflection. More generally, *Bidirectional Reflectance Function (BRDF)* describes surface reflection where the appearance depends on illumination direction and viewing direction. In inverse rendering, appearance is commonly factorized as view-independent diffuse albedo, view-dependent specular BRDF, normals and light visibility for all incoming directions [ZSD*21].

2.1.4. Camera

We also need to model the sensor that collects the input data from which we seek a reconstruction. Physically, a camera sensor collects incoming light rays (photons) and translates them into digital signals along c channels on an image grid $\mathbf{I} \in \mathbb{R}^{h\times w\times c}$ of height h × width w many pixels.

Camera Models. Camera models are described by their intrinsics. The pinhole camera model is the most widely used model; it parametrizes a perspective projection by a focal length $f \in \mathbb{R}$ and the location of the camera center $c_x, c_y \in \mathbb{R}$. Then 3D point $\mathbf{x} = (x,y,z) \in \mathbb{R}^3$ projects to the 2D point $(u,v) \in \mathbb{R}^2$ in the image plane as $u = f\frac{x-c_x}{z}$ and $v = f\frac{y-c_y}{z}$. Other models like weak perspective or orthographic camera models are also sometimes used. Due to lens distortions, real-world cameras do not follow a simple camera model exactly, which needs to be accounted for by the model if very high fidelity is desired. The extrinsic placement of the camera in the world can be described by its position, or *translation* $\mathbf{t} \in \mathbb{R}^3$, and its orientation, or *rotation* $\mathbf{R} \in SO(3)$.

Camera Types. Conventional cameras are RGB cameras that record the red, green, and blue (RGB) colors. Formally, a pixel (u,v) of an RGB image $\mathbf{I}_t$ contains the RGB color $\mathbf{c} \in [0,1]^3$

that was captured at time t. *Event cameras* are a new camera type that is, as of this writing, rarely but increasingly used [ILBH*11, GDO*22]. They follow the same camera models as conventional RGB cameras, but they record asynchronous per-pixel brightness changes instead of synchronous 2D images. Specifically, an event camera outputs a discrete stream of asynchronous events. An event (u, v, t, p) signifies that the brightness at pixel (u, v) has changed at time t by more than some threshold since the last event at that pixel. If the brightness has increased, the polarity p is $+1$, and -1 if it has decreased. The practical advantages over standard RGB cameras are a high dynamic range and microsecond temporal resolution of events. Hence, they allow to capture very fast movements with virtually no motion blur.

Monocular vs. Multi-View. A *multi-view* recording is a set of images where each dynamic deformation state is captured by more than one camera. A *monocular* recording is a recording of a dynamic scene that is not multi-view. There are two important cases: (1) a temporal sequence is a recording of a single scene across time, *e.g.* a typical video; and (2) an image collection is a set of images where each image captures a different scene (not just a different deformation state), *e.g.* Internet image search results for "people".

2.2. Problem Setting: What We Aim to Achieve

Our problem setting takes as input monocular images from a standard RGB camera or an event camera. We then seek a 3D reconstruction g_t for each point t in time. Optionally, the appearance a_t may also be reconstructed. In the case of a temporal sequence, the reconstruction is called a *4D reconstruction*. When the geometries $\{g_t\}_t$ are deformed states of a single template g, we say that the reconstructions are *in correspondence*. Instead of a single template, it is also possible to split a longer sequence into shorter pieces with their own templates (called *keyframes*).

2.3. Parametrization: Representing the Solution

In this section, we discuss how to parametrize the functions discussed in Sec. 2.1 such that they can be implemented and become tractable to compute. For example, while restricted deformation types like rigid, affine, or articulated deformations directly imply a trivial parametrization via their small set of parameters, non-rigid deformations require, *a priori*, an infinite number of parameters (one offset vector per point on the geometry). We therefore need to design approximate parametrizations that are finitely parametrized via parameters θ while still being sufficiently expressive. Parametrizations act as hard priors, *i.e.* they impose their assumptions as hard constraints. In the later Sec. 2.7 on soft priors, we will also discuss priors that can be encouraged as soft constraints.

2.3.1. Geometry Parametrizations

In the past, geometry used to be parametrized primarily by point clouds and meshes (which provide samples of the surface indicator function), and by voxel grids (which are the natural discretization of volumetric geometry functions). Since such samples of the surface indicator function (which is an implicit function) allow direct access to the surface similar to a UV map, such *parametrizations* are called *explicit*. In recent years, several concurrent works [PFS*19, MON*19, CZ19, MPJ*19] introduced global, coordinate-based multi-layer perceptron (MLP)s as geometry parametrization, both for surfaces [PFS*19] and volumes [MON*19]. Such neural parametrizations are common in Neural Radiance Field (NeRF)-like works [MST*20, WLL*21]. Since neural parametrizations typically do not provide direct access to the surface, they are called *implicit parametrizations* in both cases. (Nonetheless, neural parametrizations can also be used for explicit surface geometry via UV mapping [GFK*18].) Extracting surface meshes from these implicit parametrizations is possible with Marching Cubes [LC87, PFS*19, MST*20]. Several differentiable variants of Marching Cubes exist, which allow to define losses on the extracted geometry and then backpropagate gradients into the implicit parametrization [LDG18, RLR*20, SGY*21].

2.3.2. Appearance Parametrizations

When parametrizing a reconstruction, geometry is usually primary and appearance is afterwards attached to it in a suitable manner, *e.g.* by defining the local appearance on each mesh vertex. We note that arbitrary properties can be similarly coupled to the geometry. In works where challenging appearance is not the focus, a time-invariant Lambertian model is the first choice due to its simplicity. It can be coupled with the geometry via a UV map. If the UV map is parametrized by an image grid with appearance parameters in each pixel, it is a classical texture map. For view-dependent effects, spherical harmonics, Fourier-like basis functions on the sphere, are popular. NeRF [MST*20] was the key work that took neural parametrizations from geometry to appearance. It uses a generic position-and-view-conditioned MLP head to regress view-dependent color volumetrically, *i.e.* anywhere in 3D space.

2.3.3. Deformation Parametrizations

This section discusses explicit deformation parametrizations. Like appearance, deformations are usually attached to the geometry. In practice, applying deformation models depends on their *direction*, which greatly influences the method design. The definitions in Sec. 2.1.2 are *backward* deformation models where, in order to determine the deformed geometry g_d, we first pick a point $\mathbf{x}$ in the deformed state, apply the deformation model d to get to the reference state, and then query g: $g_d(\mathbf{x}) = g(d(\mathbf{x}))$. This warps the deformed state into the template state, which is common for volumetric representations (*e.g.* in ray tracing). If we instead turn the composition around and first query g at reference point $\mathbf{x}$ and then deform the resulting point to the deformed state, we obtain a *forward* model: $g_d(d^{-1}(\mathbf{x})) = g(\mathbf{x})$. Such models are useful for surface representations (rasterization), where one first selects a point $\mathbf{x}$ on the reference surface g and then deforms it to $d^{-1}(\mathbf{x})$. We first discuss forward models in the following, and end with backwards models for volumes.

Physics Simulation. The most accurate way to model deformations is by imposing the true physical laws that govern an object's behavior. Most methods that use physics as a hard constraint [KTE*22] are based on the Finite Element Method (FEM) from mechanical engineering. In FEM, a surface or volume is represented as a set of triangular or tetrahedral elements connected at nodes. For each element, the mass $\mathbf{M}$, stiffness $\mathbf{K}$, and the damping $\mathbf{D}$ matrices are separately built to capture their physical properties, like Young's modulus, Poisson's ratio and shear modulus. One can spatially discretize the PDE (15) to obtain an ODE in time, allowing for numerical simulation. Then, the full dynamical behavior of the object that describes the unknown vertex displacement $\mathbf{u}$ is given by:

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{D}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{f}. \tag{16}$$

Despite being ideal in principle, physics simulation is difficult to model completely and to implement, and is computationally expensive. Thus, the vast majority of 3D reconstruction works use non-physical approximations, which we discuss next.

Template Offsets. A simple deformation model consists of per-vertex offsets of a template, which is particularly popular due to recent methods trained on general image-collections [KTEM18]. As per-vertex offsets are severely underconstrained, they are often combined with soft deformation priors; see Sec. 2.7.

Skinning. As deformations tend to be spatially smooth, it is common to *skin* a detailed template geometry to a coarser graph embedded in 3D (whose parameters are thus the deformation parameters) by specifying skinning weights. When deforming the coarse graph, its deformations are transferred to the detailed template by interpolating according to these weights. Embedded graphs [SSP07] are common for general objects and skeleton skinning (itself based on a kinematic chain) is common for category-specific models [LMR*15, RTB17, ZKJB17].

Linear Subspace Models. Instead of deforming a single template, linear subspace models linearly combine a limited number of basis deformations to obtain the deformed geometry. The coefficients of this combination are often globally constant across space. This kind of parametrization is common in NRSfM [BHB00]. Similar to shape space, the low-rank assumption can equivalently be made on the trajectory or force space. Linear subspaces are sometimes used for the underlying skeletons in skinned models [LMR*15, RTB17, ZKJB17]. Models that parametrize a low-dimensional space, especially by linear combinations of some (usually fixed) basis, are called *parametric models*.

3D Morphable Models (3DMMs). More sophisticated versions of linear models where the basis is built from collections of 3D scans via statistical methods (*e.g.* PCA) are called 3DMMs. They are, thus, category-specific and most popular for faces. They factorize deformations into independent identity and expression parameters. In addition, they often include appearance. Most morphable models work on a mesh geometry and are created via PCA, which is linear in the instance-specific coefficients. Recently, there are attempts to learn volumetric neural morphable models that are non-linear in the instance-specific latent code; see Sec. 3.3.2.

Volumetric Deformations. Implicit geometry parametrizations tend to use volumetric backward deformation models to avoid the need for directly accessing the surface. The earlier work Neural Volumes [LSS*19] uses an explicit mixture of affine warps, while recent neural-rendering methods use an MLP parametrization. In the fully non-rigid case, such an MLP can output an offset per point in 3D space [TTG*21], while more articulated deformations benefit from an $SE(3)$ output [PSB*21]. Since these are fully unregularized (up to the smoothness of the MLP and soft deformation priors), there are also first attempts to extend skinning to the volumetric case [CZB*21, YVN*22].

2.3.4. Camera Parametrizations

The definitions of camera intrinsics and extrinsics imply direct, natural parametrizations. For example, a 3D translation can be trivially parametrized by a 3D vector. Solely camera rotations, which are elements of the 3D rotation group $SO(3)$, inherently cannot have a natural (smooth, unique, without boundary) parametrization (this is ultimately due to the universal cover of $SO(3)$ being a *double* cover by $SU(2)$). Common parametrizations are Euler angles, axis-angle, quaternions, and rotation matrices. We refer to Zhou *et al.* [ZBL*19] for details. The distorted ray directions caused by lens distortions can be parametrized well by correctives following, for example, the Brown-Conrady model [Bro66].

2.3.5. Large-Scale Image Collections

While it is tractable for temporal sequences and small-scale image collections [YHL*22] to directly optimize for the reconstruction parameters of the scene, this becomes impractical for large-scale image collections with thousands of images. Instead, the scene-specific parameters θ_s are output by a meta-reconstruction function $\theta_s = f_\theta(s)$ (called *data-driven prior*) that accumulates generalizable reconstruction knowledge about the image collection it is fit on. In practice, this data-driven prior is typically a neural network (specifically, a CNN in the case of input images) that regresses the scene-specific parameters of its input scene.

2.4. Rendering: Connecting 3D and 2D

In the reconstruction setting, we are provided with 2D input data which we need to relate to the 3D model. To that end, rendering is crucial as it allows us to extract 2D information from the 3D model. Given the scene decomposition consisting of lights, material, and (deformed) geometry, a virtual camera generates a 2D observation of the 3D world in the rendering process. Rendering is the computational model of the physical light-collecting process of a camera.

Rendering. Works on reconstruction employ a small set of standard rendering techniques. Explicit geometry parametrizations like meshes or point clouds (see Sec. 2.3.1) are typically rendered using *rasterization*, which projects each geometric primitive (*e.g.* triangle or point) using a virtual camera. If instead an implicit or volumetric geometry parametrization like an MLP or a voxel grid is used, *ray tracing* is typically applied. For each pixel of the virtual camera, it traces a ray into the scene, trying to hit geometry. For surface geometry functions, surface rendering can be used [NMOG20, SZW19], which picks the first surface along the ray as the point to be rendered, while volumetric geometry functions can use volume rendering, which accumulates geometry and appearance along the ray [MST*20, LSS*19].

Inverse Rendering is the inverse operation of rendering, *i.e.* recovering the intrinsic components (geometry, material, illumination, and deformations) of a 3D or 4D scene from images. To that end, we can exploit (forward) rendering for *analysis by synthesis*, where we obtain the 3D or 4D scene reconstruction (analysis) *by ensuring* *that it can render (synthesize) the 2D input*. (This STAR also covers methods that primarily use 3D supervision at training time and hence do not follow the analysis-by-synthesis paradigm.)

Differentiable Rendering. Rendering is naively not differentiable and hence prevents gradients from propagating from the image loss to the 3D model. In the simple case of point-based rendering, bilinear interpolation of the input image provides gradients to each 3D point [TZK*17]. Several works introduce methods that make mesh rasterization differentiable [LB14, KUH18, LLCL19]. For ray tracing, differentiable surface rendering is challenging because determining the surface intersection is not naturally differentiable [NMOG20, SZW19]. However, differentiable rendering of a volumetric scene is rather straightforward with volumetric rendering because no surface needs to be determined [LSS*19, MST*20], which also provides a workaround for differentiable surface rendering [WLL*21].

2.5. Data Terms: Ensuring Consistency With the Input

Now, we turn to the inverse-problem aspect of the reconstruction problem. We require data terms that fit the model to the input data by encouraging consistency between the reconstruction and the input. When provided as input, consistency is usually easy to obtain with camera extrinsics and intrinsics, a template geometry, boundary points, timestamps, or a texture: we simply set the parameters of our geometry parametrization, for example, to the template geometry. The consistency is therefore “hard” in these cases. Other inputs, which lack such a nice correspondence to the parametrization, are more difficult and typically consistency is merely “soft”, *i.e.* encouraged (but not enforced) via losses.

2.5.1. Common Data Terms

Since supervision most often happens via 2D input data, we need to render our model into 2D and then compare to the input data. In the case of RGB image input, typical photometric losses are ℓ_1 or ℓ_2 losses, and, in recent years, perceptual losses like LPIPS [ZIE*18]. Similar to these appearance-focused losses, 2D object segmentations are usually easy to obtain from the model geometry and can then be compared to input segmentations masks, which tends to help with coarser mismatches. Correspondences across time can also be extracted from the deformation parametrization and then be fitted to dense 2D input correspondences (optical flow) or 3D input correspondences (scene flow [ZXLK21]). Sparse 2D correspondences from feature point tracking (*e.g.* via SIFT [Low04]) are sometimes used too, as they help with reconstructing large deformations. When correspondences across images—and not images themselves—are the input to the method, the latter falls into the category Non-Rigid Structure-from-Motion (NRSfM). In analytical Shape-from-Template (SfT) methods, matches between the template and input image are provided instead of an RGB image. In the case of image collections, we might be given certain model parameters as input (*e.g.* morphable model parameters). The estimated deformation parameters and the input parameters can then be compared in a similarity loss.

2.5.2. Other Data Terms

Beyond appearance-based matching, some recent methods exploit *2D-3D consistency of learned features*. Thus, ViSER [YSJ*21b] learns them from scratch, BANMo [YVN*22] uses Continuous Surface Embeddings [NNS*20], and LASSIE [YHL*22] uses DINO [CTM*21] features. A rendering loss encourages consistency between the features attached to the geometry and the image features. Unlike appearance, these features can more readily incorporate local and global context, and hence provide more information about larger-scale mismatches. Typically used in a generative setting, *3D-aware GANs* use a 3D representation in the generator to render 2D images. A 2D discriminator then encourages those images to resemble the distribution of some given set of input images. Since this imposes consistency with input data in a looser fashion, it leaves the generator more freedom to hallucinate finer details that look plausible, instead of having to reconstruct the input exactly. To apply such GANs to reconstruction, they are first trained in a generative manner and then need to be inverted at test time, *i.e.* the right latent code for some input image needs to be determined. See *2D Supervision* in Sec. 3.3.2 for more details.

2.6. Challenges: What Makes the Problem Difficult

Unfortunately, there are challenges on multiple levels when trying to find a 3D reconstruction. In this section, we discuss a variety of them and mention some potential solutions. The next section focuses on the main challenge: the underconstrained nature of the 3D reconstruction problem and priors to tackle it.

2.6.1. Inherent Challenges

Several issues are inherent to the problem formulation and cannot be solved by any amount of data.

Challenge: Occlusions. Especially in the monocular setting, an object may self-occlude, *e.g.* due to the movement of human body parts and folding of cloths, or become occluded due to an external object. We thus have no input information about its current state and the data terms are unavailable, *a priori* preventing reconstruction. This is a root challenge in monocular reconstruction.

Solution: Regularization. Soft priors (Sec. 2.7) are used to fill-in missing information. We note that many methods do not explicitly consider occlusions and instead apply the same prior to occluded and visible regions.

Depth Ambiguity. Another root challenge is the lack of depth when using monocular visual measurements. All points on the (optical) ray are projected to the same image point, leading to depth ambiguities: Different 3D geometries can lead to the same 2D projection

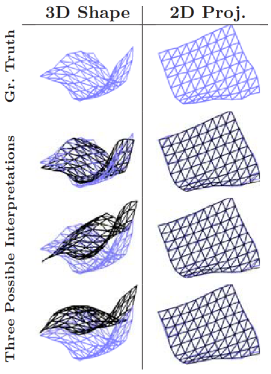

Figure 4: *2D-3D depth ambiguity. Image adapted from [MNPF10].*

in image space, as Fig. 4 shows. Several special cases arise from this, which we discuss next.

Geometry-Appearance Ambiguity. Correctly attributing fine-scale image details to geometry versus appearance (*e.g.* in the case of fine wrinkles and textures) is *a priori* ambiguous.

View Dependence. Arbitrary view-dependent appearance makes it possible to attribute the image formation to almost any geometry with view-dependent appearance.
Few Degrees of Freedom (DoFs) for View Dependence. The DoFs are typically limited by using only spherical harmonics up to degree three or using a very small MLP for NeRF-style view dependence.

Focal Length. In image collections, the actual focal length typically varies for each image. However, a single image is insufficient to estimate the focal length.
Fix to Arbitrary Value. In practice, it is common to approximate the focal length by a fixed value across all images.

2.6.2. Parametrization Challenges

Topology Change. Topology changes occur when a surface starts merging or splitting apart. They are difficult to handle because they need to be detected and then accounted for in the parametrization.
Discarding Correspondences. Current methods that handle topology changes do so by not having a single, consistent geometry parametrization across time, thereby discarding correspondences.

Discretizing Losses. Loss functions, consisting of data terms or priors, are in most cases easier and more intuitive to formulate in a continuous manner. Discretizing such continuous formulations onto discrete parametrizations (like meshes) is non-trivial, and one continuous formulation may give rise to different discretizations with different optimization behavior and formal guarantees.

Camera-Rigid-Motion Ambiguity. Without static background, the rigid motions of the camera and object are ambiguous.
Assume Static Camera or Arbitrarily Factorize. Small camera motion can be modeled as a rigid transform of the object under a static camera, especially for SfT. Some methods, in particular in NRSfM, separately account for the camera movement, and output the camera rotation and translation.

Identity-Deformation Ambiguity. When reconstructing faces or humans, it is often desired to factorize the deformations of the template geometry into identity-specific (invariant per person) and pose-specific (varying over time) components. However, image collections often contain only one image per person, making such a factorization *a priori* ambiguous.

2.6.3. Data Acquisition Challenges

Background. Static background is visible when recording.
Ignoring, Partial Modeling, Full Modeling. Most methods remove it at the input level via background subtraction (for static cameras) or image segmentation. Especially in the case of image collections, the background is sometimes kept in the input but then the reconstruction either ignores it or reconstructs it only badly [WRV20]. Recently, a handful of NeRF-based methods properly reconstruct the static background as well, see Sec. 3.1.3.

Motion Blur. When recording a fast-moving scene or moving a camera quickly, a pixel may collect color from different points of the scene within the short time frame when the sensor is active for the current frame. This leads to so-called *motion blur*.
Filtering. Motion blur is difficult to account for in a model and is hence seldom modeled. Instead, blurry images tend to be discarded or heuristically de-blurred during pre-processing.

Lens Distortions. Lens distortions can be decently well estimated for a temporal sequence, while image collections are too severely underconstrained, similar to the focal length ambiguity.
Undistortion. In most works that consider lens distortions at all, the forward models are applied to obtain an undistorted pinhole-camera image, although some works instead optimize for the corrected ray directions of the distorted image [PSB*21].

Noise. Noise in the input (*e.g.* RGB images, camera parameters, drifting correspondences) is, by its nature, not easy to detect. Even small noise can have negative impacts on the optimization.
Correcting Input Estimates. Some estimated input parameters θ_{est}, especially camera extrinsics and intrinsics, tend to be slightly incorrect (noisy) in practice. If this noise is too severe, it is possible to optimize for corrective parameter offsets $\Delta\theta$: $\theta_{\text{correct}} = \theta_{\text{est}} + \Delta\theta$, where $\Delta\theta$ is usually kept small via an ℓ_1 or ℓ_2 loss.
Robust Losses. Some losses are more robust to input outliers, *e.g.* a Huber loss or an ℓ_1 loss more so than an ℓ_2 loss.

On-Camera Processing. On a very practical level, modern cameras process images, which might lead to undesirable effects (*e.g.* automatic white balancing or varying gamma correction, as well as missing color calibration for image collections can lead to varying measurement results of the exact same real-world color).
Ignoring It, Modeling It or Turning It Off. Oftentimes, the parameters of these operations are not accessible to the end user, and hence are ignored. They can also be estimated afterwards if these kinds of processing cannot be turned off when recording.

2.7. Soft Priors: Making the Problem Tractable

A lot of information is lost during the image formation process because, at any time step, we only obtain a *visual* measurement under *one* viewing angle of any *visible* surface point. We thus need to fill-in this lost information. In addition, we need to prevent undesirable local minima to stabilize the optimization in practice. As with other inverse problems, we therefore seek to constrain/regularize the solution space of the shape reconstruction with prior assumptions, ideally forcing the existence and uniqueness of a solution. We note that Sec. 2.3 discusses hard priors.

2.7.1. Geometry Soft Priors

Geometry priors are solely spatial, *i.e.* they only act on a geometry by itself, regardless of whether it was obtained through deforming some reference geometry or not. Typical priors include spatial smoothness, where, for example, a Laplacian loss or a loss on the normals of the geometry encourage locally smooth geometry, and parametrization-specific priors that discourage local minima, like a loss encouraging mesh edges to be short. Some methods may exploit symmetry constraints.

2.7.2. Deformation Soft Priors

Most deformation priors act on the final shape by introducing one or more reference geometries with respect to which they regularize the current one. Alternatively, they can rely on parametric models and regularize their parameters.

Metric-Based Priors. Many spatial deformation priors approximate the underlying physical properties of non-rigid objects. They are defined locally and follow from the measures of deformation (see Sec. 2.1.2). Physically plausible deformations are assumed to preserve different metric quantities such as lengths, angles, and areas on the surface of the geometry. The measures are defined on a deformed geometry g_d, with respect to a reference geometry g. Ideally, the reference geometry should be a physical rest pose so that the deformations are not just geometrically but also physically meaningful. Then with the definitions from Sec. 2.1.2 and Fig. 3,

- *Isometric deformation maps* preserve the geodesic distance between any two points on the surface (*e.g.* consider paper):

$$G(\mathbf{x}(\mathbf{m},t)) = G(\mathbf{x}(\mathbf{m},0)) \tag{17}$$

  It only allows surface bending, but not stretching or shearing. A simpler alternative is *inextensibility*, which preserves the *Euclidean* distances instead. For real-world extensible objects, *quasi-isometry* prevents large stretching or shrinking and can be implemented as As-Rigid-As-Possible (ARAP) prior [SA07].
- *Conformal maps* preserve local angles on the surface:

$$G(\mathbf{x}(\mathbf{m},t)) = \lambda(\mathbf{m})G(\mathbf{x}(\mathbf{m},0)) \tag{18}$$

  Conformality is weaker than isometry and allows for stretching (*e.g.* an expanding balloon).
- *Equiareal maps* preserve area on the surface and lead to:

$$\det(G(\mathbf{x}(\mathbf{m},t))) = \det(G(\mathbf{x}(\mathbf{m},0))) \tag{19}$$

  Isometry is equivalent to conformality and equiareality together.

Other Reference-Based Priors. In addition, we may also favor stricter closeness to the template. For example, we may encourage the template offsets to be small, or, in the case of skeleton-based deformations, the angles of the joints to stay close to the rest pose or within a certain range. Furthermore, we can encourage the coefficients or latent codes of a parametric model to be close to zero.

Temporal Priors. Unlike single images, videos provide an additional temporal dimension that can be leveraged as a prior. Assuming that images are sampled at high enough frame rates, deformation states that are temporally close are, in general, similar to each other. We can impose this prior knowledge about temporal smoothness using similarity losses between the deformations of the time steps in question. Alternatively, when reconstructing a temporal sequence of multiple time steps, it can be useful to optimize in a sequential manner, starting with reconstructing $t = 0$ and then continuing step-wise for $t > 0$. In particular, the deformation parameters $\theta_d^t \subset \theta$ at time t are often initialized from the previous time step: $\theta_d^t = \theta_d^{t-1}$, which is usually referred to as *tracking*.

2.7.3. Appearance Soft Priors

In the case of texture-less surfaces, a smoother change in appearance or more explicit priors about lighting and reflectance maybe employed to aid reconstruction using shading cues. Nonetheless, soft priors are only rarely applied for appearance. 3DMMs [BV99] and similar statistical models can encourage the appearance parameters to stay close to the estimated prior parameter distribution, which is often assumed to be normal distributed.

2.8. Optimization: Finding the Right Parameters

Once we have set up a solution parametrization with a set of parameters θ and a loss function $\mathcal{L}(\theta)$, containing data terms and priors, we can finally determine the best set of parameters θ^* as the solution to the 3D reconstruction problem:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\theta). \tag{20}$$

There is a wide variety of optimization techniques that is used in the literature for this problem that is virtually always highly non-convex. While neural methods are almost exclusively optimized via gradient-based techniques (using $\frac{\partial \mathcal{L}}{\partial \theta_i}$ for the i-th parameter) that start from an initial guess θ_{init}, other methods also employ gradient-free optimization (such as simulated annealing, particle swarm optimization or evolutionary policies). A detailed discussion of these techniques, however, is outside the scope of this section.

3. State-of-the-Art Methods

The main axis along which we organize our discussion is the object category that is to be reconstructed. After discussing methods for monocular 3D reconstruction of general objects (Sec. 3.1), we describe the state of the art of methods specialized for the human body (Sec. 3.2), faces (Sec. 3.3), hands (Sec. 3.4), and animals (Sec. 3.5). We discuss methods using event cameras in Sec. 4.

3.1. General Objects

We first discuss the established fields of SfT (Sec. 3.1.1) and NRSfM (Sec. 3.1.2) before moving on to few-scene reconstruction methods that rely neither on template nor correspondences as their core assumption. We split these into NeRF-like neural methods (Sec. 3.1.3) and others (Sec. 3.1.4). We finally turn to data-driven approaches (Sec. 3.1.5) that work on large-scale image collections.

3.1.1. Shape from Template (SfT)

Shape from Template (SfT), or template-based reconstruction, comprises monocular non-rigid 3D reconstruction methods that assume a single static shape or *template* is given as a prior. It has been an active research area for two decades [SLF07, SF10]. The name SfT (not to be confused with Shape from Texture) became common after 2015 due to the eponymous work of Bartoli *et al.* [BGC*15]. We next discuss templates, solving strategies and deformation priors in SfT before describing the state of the art in detail.

Template. Given a template in a reference configuration and a calibrated camera, SfT aims to reconstruct the shape of a deformable

Object Type	Solving Strategy	Data Term	Deformation Prior	Temporal Coherency
curve: [GPCB20] volumetric: [PPBC15, FJCPP*18, YRCA15] thin-shell: [KTE*22, CPPFJ*21, SGTS19, FJCPP*21]	analytical: [CPPFJ*21, CPPFJ*19, CPBC16, BGC*15] energy-based: [KTE*22, ÖB17, MH17, YRCA15] neural (object-specific): [FJCPP*18, SGTS19, PAP*18, GSVS18] neural (generic): [FJCPP*21, SGTS19]	template-image warp: [CPPFJ*21, NPJ*15] per-pixel intensity: [KTE*22] per-vertex intensity: [ÖB17, YRCA15] shading cue: [LYYA*16] pre-trained: [SGTS19, FJCPP*21] surface micro-structure: [HXR*18]	isometry: [CPPFJ*21, SGTS19, CPBC16] conformality: [BGC*15] equiareality: [CPPFJ*19] elasticity: [KTE*22, ÖB17, MH17] ARAP: [FJPCP*21, YRCA15] Laplacian: [NÖF15] low-rank: [TTZ*20]	present: [KTE*22, YRCA15] not present: [SGTS19, CPPFJ*21, FJPCP*21]

Table 1: *Overview and classification of Shape-from-Template methods.*

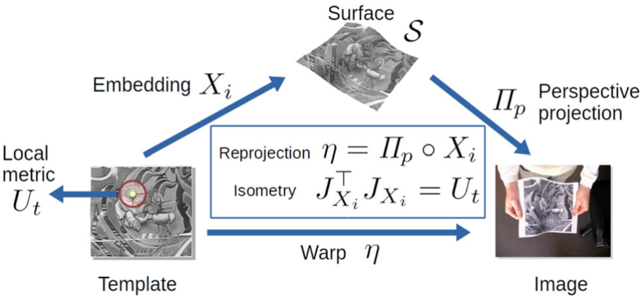
Surface $\mathcal{S}$
Embedding X_i
Π_p Perspective projection
Local metric U_t
Reprojection $\eta = \Pi_p \circ X_i$
Isometry $J_{X_i}^\top J_{X_i} = U_t$
Warp η
Template
Image

Figure 5: *Differential isometric SfT [CPPFJ*21, CPBC16, BGC*15] solves for the reconstruction embedding X_i, given the warp η (which aligns the template with the input image). Image adapted from [CPPFJ*21].*

object in every frame of a video sequence observing the object. The template often corresponds to the first frame of the sequence though it is not always a strict requirement. The template can be used as the initial state of a physics simulator [KTE*22], to obtain 3D-2D registration as a basis for reconstruction [CPPFJ*21], and to encode prior knowledge in neural network weights [SGTS19, FJPCP*21].

Solution Strategies. As shown in Tab. 1, SfT methods can be classified as *energy-based*, *analytical*, and *neural-based* approaches. Energy-based methods [KTE*22, ÖB17, MH17, YRCA15, BBH14, MBH15] define a non-convex cost function with photometric consistency as the data term and deformation priors acting as the regularization term. The energy is typically minimized using iterative optimization methods [KB15, Mor78]. One major challenge for this method class is that the energy landscape is often nonlinear, and the algorithm can potentially converge to erroneous local minima. Therefore, careful initialization is required, and the template is often used as an initial shape [KTE*22, ÖB17]; alternatively, the solution from the previous frame may also be used as in [YRCA15]. Analytical methods [CPPFJ*21, CPPFJ*19, CPBC16, BGC*15, FRA11] formulate re-projection and deformation constraints as PDEs and provide well-posed analytic solutions in a single step. However, these do not match the accuracy of energy-based methods and require a refinement of reconstructions using iterative methods. Recently, neural methods [FJPCP*21, FJCPP*18, SGTS19, PAP*18, GSVS18, TTZ*20] have been used to learn image to 3D shape mappings by training deep networks on datasets of deforming sequences. Since, at test time, 3D reconstructions are obtained simply by a single feed-forward pass, they usually achieve a higher runtime performance compared to energy-based approaches. However, they tend to be texture- and template-specific and often have difficulty generalizing to unseen shapes.

Deformation Priors. SfT methods can also be classified according to the type of deformation priors. Strong ones (*e.g.* isometry) have been extensively studied [BGC*15], whereas weaker but more accurate elastic priors are becoming increasingly popular [KTE*22, CPPFJ*19, MH17]. In early works [CPBC16, BGC*15], registration between template and input image, along with their differential structures and isometric constraints, delivered well-posed problems with unique solutions. More recently, neural networks [SGTS19, GSVS18, NGM*21] have been used to favor isometry instead of enforcing it. For extensible surfaces such as balloons, conformal geometric prior have been similarly used to obtain families of solutions [BGC*15]. Casillas-Perez *et al.* [CPPFJ*19] provide a theoretical framework for equiareal SfT using Monge's theory for solving first-order nonlinear PDE and show results on stretched fabrics. Parashar *et al.* [PPB19] use Cartan's theory of connections and moving frames, that offers a generic solution to all local (isometric, conformal and equiareal) deformation models. While these geometric priors are only approximate, physically exact stretching and bending priors can be derived from the continuum mechanics of elastic objects. The approach of Malti *et al.* [MBH15] relies on linear elasticity to minimize stretching energy under reprojection boundary conditions, which was later extended to constrain the set of spatial forces to be sparse [MH17]. Similarly, Özgur *et al.* [ÖB17] specify stiffness parameters describing the stretching and bending behavior of elastic objects, whereas another method uses isotropic material elasticity (Saint-Venant Kirchhoff model) [HC17]. To model non-linear and anisotropic behaviors of challenging cloth deformations, ϕ-SfT [KTE*22] imposes the elastic model of [WOR11] as deformation prior.

State-of-the-Art Methods. Given the *warp* relating template to the input image and their differentials, analytical methods [CPPFJ*21, CPBC16, BGC*15] formulate the SfT problem in terms of a system of non-linear first-order PDEs, as shown in Fig. 5. These equations depend on the unknown reconstruction embedding X_i, uniquely defined with the depth function ρ, given the warp η and the template's local metric U_t. Bartoli *et al.* [BGC*15] directly solve for depth ρ and its derivatives ∇ρ as independent variables in the isometric SfT system, not related via differentiation, leading to the non-holonomic solution. Extending this, Chhatkuli *et al.* [CPBC16] propose to use the non-holonomic depth's gradi-

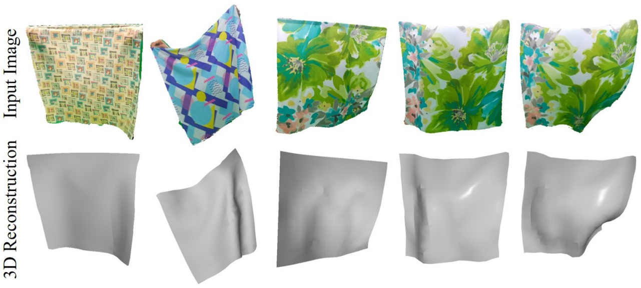

Figure 6: ϕ-*SfT [KTE*22] explicitly simulates the physical fold formation process. It can thus handle even challenging local folds. Image adapted from [KTE*22].*

ent to recover the surface via integration. This strategy is significantly more stable and robust to errors in the warp. Alternatively, Casillas *et al.* [CPPFJ*21] propose *isowarp* to improve the warp for the analytic depth solutions [BGC*15]. They define a set of warp constraints from the 3D isometry equations, and the resulting warp representation improves the accuracy of reconstructions.

A recent real-time SfT approach by Fuentes-Jimenez *et al.* [FJPCP*21], *i.e.* RRNet-DCT, relies on deep neural networks. Its architecture has two neural networks: A segmentation module for pixel-based detection of the template and a registration-reconstruction module to perform SfT. RRNet-DCT is texture-agnostic as it adapts to new texture maps at runtime compared to the authors' earlier texture-specific method, DeepSfT [FJPCP*21]. Being an object-specific method that encodes the template into the neural network weights, it is highly accurate, unlike earlier object-generic methods such as IsMoGAN [SGTS19]. However, both DeepSfT and IsMo-GAN are less generic methods than energy-based methods.

In contrast to the *wide-baseline* analytical and neural methods, the recent *short-baseline* method ϕ-SfT by Kairanda *et al.* [KTE*22] leverages the temporal consistency across frames. ϕ-SfT accounts for 2D observations through physical simulations of forces and material properties. They use a differentiable physics simulator [LLK19] to regularize the surface evolution and to optimize the forces and material elastic properties such as bending coefficients, stretching stiffness and density. Following an analysis-by-synthesis approach, a differentiable renderer is employed to minimize the dense reprojection error between the estimated 3D states and the input images; the deformation parameters are recovered by adaptive gradient-based optimization. Compared to earlier analysis-by-synthesis solutions with per-vertex photometric costs [YRCA15, LYYA*16], ϕ-SfT's per-pixel approach uses the full information in the high-resolution texture map, leading to accurate reconstruction of challenging local folds; see Fig. 6.

Datasets. SfT methods require reference templates and image sequences as part of the dataset. The template and the respective texture map are generally acquired with SfM [YRCA15] or an RGB-D camera [KTE*22]. Most works also evaluate on real and synthetic datasets that satisfy the assumptions on deformation types of the respective methods. We list the real datasets with the most recent ones first: ϕ-*SfT* [KTE*22]; *t-shirt* and *balloon* and *sock* [CPPFJ*19]; *zooming* and *can* [CPBC16]; *face*, *bobby* and *pig* [YRCA15]; *cushion*, *balloon* and [BGC*15]; *woggle*, *sponge* and *arm* [PPBC15]; *balloon*, *spandex*, *redchecker* and *cap* [MHBK13, MBH15]; *t-shirt* and *paper* [VSFU12]; and *face* [VWB*12]. Besides real sequences, a few methods also evaluate on synthetic datasets whose geometries are generated with physics simulation. Learning-based works [GSVS18, PAP*18] often train neural networks using lightweight, synthetic training datasets. They incorporate various deformations, textures, illuminations and camera poses to ensure generalizability to unseen images.

Open Challenges. SfT has been successfully applied in the medical domain (*e.g.* to register a preoperative 3D liver model to a laparoscopy image [EÖC*20, CBBC16, KÖR*17]); however, practical applications are still limited, and we list the reasons for the same. A set of problems not yet attempted in the field include: background reconstruction, changing object topology, multiple deformable objects and severe self-collisions. Next, it is common to evaluate SfT on datasets with smooth deformations (*e.g.* the *t-shirt* and *paper* sequences [VSFU12]). The physics-based ϕ-*SfT* approach [KTE*22] supports challenging local folds but fails to capture small and frequent wrinkles. Implicit surface representations have not yet been studied in the context of the classical SfT problem, but we believe they have potential as in many other subfields. Non-learning methods use triangular meshes with a fixed resolution, while learning-based SfT techniques have fixed output sizes. Despite offering unique and closed-form solutions—as registration with a template is fundamental to analytical SfT—errors in warps propagate to 3D and limit the reconstruction accuracy. SfT methods commonly operate on individual images, and although they provide 3D correspondences with a template, the reconstructions can suffer from frame-to-frame jitter. Deviating from this, a few works [YRCA15, HXR*18] employ an explicit temporal regularization term and ϕ-SfT outputs temporally smooth surfaces owing to simulation. Besides, exploring joint optimization over multiple frames is promising and tractable for SfT due to advances in GPUs.

3.1.2. Non-Rigid Structure from Motion (NRSfM)

Whereas SfT uses the information present in a single image to deform the template, NRSfM relies on motion and deformation cues for 3D recovery of deformable surfaces [BHB00] and is more generally applicable than SfT. The input to NRSfM are 2D point tracks across multiple images, also called *measurements* or *measurement matrices*, and the output is a set of per-view camera-object poses and 3D shapes. This section, similarly to the entire STAR, focuses on dense NRSfM methods, which operate on (per-pixel) densely tracked 2D points. During dense point tracking with optical flow or video registration methods [GRA13b], a single keyframe is selected, and the 3D points visible in it are tracked across all remaining views and subsequently reconstructed. While sparse NRSfM approaches treat every input point independently, dense approaches assume that the observed surfaces are spatially coherent.

NRSfM uses only *weak* prior assumptions about the observed motions and deformations and no 3D priors. Significant progress was achieved in comprehending and solving this classic ill-posed 3D computer vision problem over the last decades [Bra05, THB08,

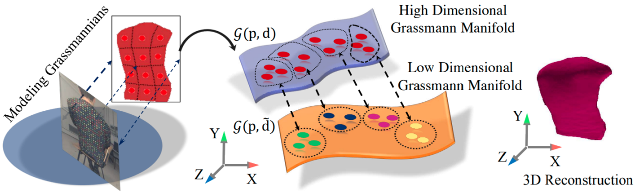
Modeling Grassmannians
$\mathcal{G}(p, d)$
High Dimensional Grassmann Manifold
Low Dimensional Grassmann Manifold
$\mathcal{G}(p, \bar{d})$
Y
Z
X
3D Reconstruction

Figure 7: *Overview of Jumping Manifolds [Kum19]. Image adapted from [Kum19].*

GM11, PDBX*12, DLH12, GRA13a, KL16, AGS17, KCDL18]. State-of-the-art and highly influential NRSfM methods at different times were Bregler *et al.* [BHB00] (the first NRSfM method), hierarchical approach [THB08], trajectory-space method [ASKK08], NRSfM with minimal prior assumptions [DLH12], variational approach [GRA13a] as well as multi-body NRSfM [KDL16].

State-of-the-Art Methods. The first dense NRSfM methods [RFA12, GRA13a] provoked many follow-up works. Most state-of-the-art methods follow (at least implicitly) the matrix factorization approach of Bregler *et al.* and the prior assumption that the deformable shapes span low-rank subspaces [BHB00].

Several works [ZHDLTL14, AGS17, KCDL18, Kum19, Kum20] were inspired by the Block Matrix Method (BMM) of Dai *et al.* [DLH12]. BMM is convex and only assumes the low-rank shape constraint; it showed that NRSfM could be solved unambiguously w.r.t. the basis unknown during optimization. The SMSR method of Ansari *et al.* [AGS17] updates the input measurement matrix by applying smooth trajectory constraints. Differently from Dai *et al.*, they use the alternating direction method of multipliers (ADMM) to optimize the objective function. Moreover, SMSR converges fast and scales well across datasets of different point sizes. The jumping manifolds (JM) approach [Kum19] is an extension of Grassmannian NRSfM (GM) [KCDL18]. Both methods follow the ideas of point clustering and unions of linear subspaces [ZHDLTL14]. JM takes into account that local surface deformations depend on point neighborhoods. It combines high and low-dimensional Grassmann manifolds for 3D reconstruction and clustering; see Fig. 7 for an overview of the method. JM currently achieves one of the lowest 3D reconstruction errors on one of the synthetic faces [VBPP05, GRA13a]. The weaknesses of GM and JM is an excessive number of parameters that need to be set compared to many other techniques requiring much fewer of them [PSF20, STG*20, GB22, WLPL22].

Sidhu *et al.* [STG*20] introduced N-NRSfM, *i.e.* the first *neural* dense NRSfM approach with a deformation model represented by a neural network. They follow the auto-decoder paradigm and assign a latent space variable to each 3D state; see Fig. 8. The deformation model of N-NRSfM provides sufficient expressiveness due to non-linearities of the MLP, and the latent space function (*i.e.* the set of per-shape latent variables) compresses the reconstructions into a lower-dimensional space. A new loss Sidhu *et al.* impose is the latent space constraint in the Fourier space that forces similar 3D shapes—observed in arbitrary frames—to have similar latent variables. It also allows to reveal periods of the input sequences.

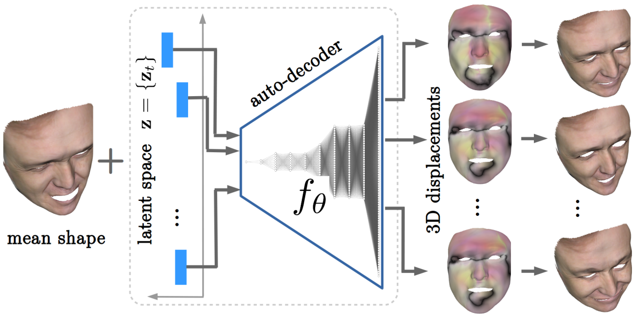

Figure 8: *Deformation model of N-NRSfM [STG*20], the first NRSfM approach with neural deformation model. Image adapted from [STG*20].*

Wang *et al.* [WLPL22] proposed a neural trajectory prior (NTP) for motion regularization in different 3D computer vision tasks, including scene flow integration and dense NRSfM. NTP relies on the smoothness bias of MLPs and imposes temporal smoothness and spatial similarity on continuous point trajectories. Similarly to N-NRSfM [STG*20] and PAUL [WL21], they use a bottleneck layer in their model, which forces the resulting trajectories to be compressible (*i.e.* to lie in a low-dimensional space).

Graßhof and Brandt combine tensor-based modeling and rank-one 3D shape basis formulation for NRSfM [GB22]. They recover 3D shapes up to an affine 3D transformation and perform a metric update if camera calibration is known. They primarily target 3D reconstruction of faces and achieve accurate results on the BU3DFE dataset [YWS*06] compared to several previous methods.

A different approach for smooth surfaces is pursued in Diff-NRSfM [PSF20]. This method assumes local surface diffeomorphism associated with specific differential properties of the 3D points. Diff-NRSfM is among the fastest methods achieving competitive performance in dense scenarios. Few works target occlusion handling or restricted camera paths in dense NRSfM [GFS17a, GJST20, SB21]. They are motivated by medical applications (*e.g.* endoscopy), in which relying on 2D matches only is not sufficient. SPVA [GFS17a] combines NRSfM with SfT for increased 3D reconstruction stability while handling inaccurate and partially corrupted dense correspondences (*e.g.* due to large external occlusions). DSPR [GJST20] extends this idea to a dynamic shape prior with multiple 3D states obtained on non-occluded parts of the input sequence. The dynamic shape prior is then used to stabilize the occluded shape parts (*e.g.* by a robotic arm), while the non-occluded regions select the most suitable 3D surface (that was previously observed) for 3D shape inpainting. Fig. 9 shows DSPR's exemplary 3D reconstructions of the heart bypass sequence [Sto12]. A proof-of-concept approach with topological shape prior [SB21] assumes that the reconstructed shape is tube-shaped, as expected in colonoscopy. It alternates between unconstrained 3D reconstruction assuming isometry and tubular parametrization upgrading the initial point clouds to tubular-shaped smooth surfaces.

Datasets. NRSfM is a severely ill-posed problem, and no single NRSfM method was shown to reconstruct sequences observing different motion and deformation types with steadily high accuracy.

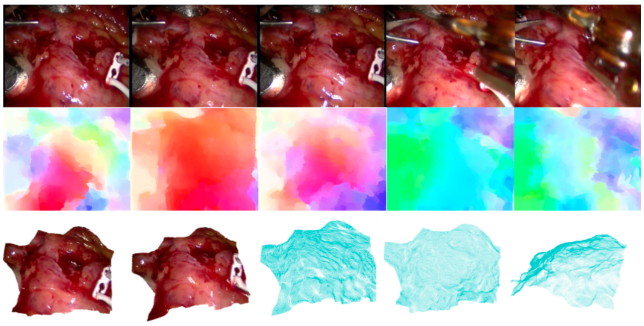

Figure 9: *DSPR's [GJST20] results (bottom) on the 2D input correspondences (middle) of the bypass surgery sequence [Sto12] (top). DSPR detects and handles the robotic arm as an occluder. Image adapted from [GJST20].*

Dense NRSfM approaches were tested on different sequences and types of non-rigid objects over the last ten years. We summarise most of them in the following by starting with the ones providing 3D ground truth: *Synthetic flag* [GRA13a, GRA13b], *synthetic flag with occlusions* [GFS17a], *synthetic faces* [VBPP05,GRA13a], actor [BHB*11, AGS17], *toss* and *pants* [WCF07, AMN18], *actor mocap* [VWB*12,GJS19,GJST20]; the sequences *t-shirt* and *paper* [VSFU12] coming with reference depth data recorded by a Kinect sensor. Widely-used sequences without ground truth are: *Face* ("Nico") [GRA13a], *back* [RFA11], *heart bypass surgery* (two sequences) [SMD*05, Sto12], *rabbit laparoscopy* [AMNCM16, GFS17b], *liver* [MSY10, GJST20] and *barn owl* [VGS16]. Note that a few NRSfM methods for dense reconstruction [AGS17, GJS19,PSF20] were also tested on the sparse (semi-dense) NRSfM Challenge 2017 dataset of Jensen and colleagues [HBAD21].

Field Specifics and Open Challenges. Despite all the progress, there remains a significant gap between the theory and practical applications of NRSfM, and several reasons for that can be named.

First, the input 2D point tracks are usually extracted from the input views by dense optical flow techniques [GRA13b, TBGS16]. Unfortunately, most NRSfM papers ignore the recent progress in optical flow estimation, even though 1) modern deep-learning-based methods [TD20] can be applied to deformable objects, and while 2) it is well known that the accuracy of NRSfM depends on the accuracy of point tracks. Many NRSfM datasets, however, provide ground-truth 2D correspondences obtained by re-projecting ground-truth 3D shapes to an image plane by a virtual camera. This allows to focus on the 3D reconstruction while delegating dense point tracking. At the same time—even if a method can accurately reconstruct a scene from accurate point tracks—it is often not known how the same approach performs on real and deteriorated 2D tracks [GRA13a, AGS17, GB22, WLPL22]. (Only several works evaluate the proposed methods on noise-contaminated ground-truth measurements [KCDL18, Kum19, PSF20, GJST20].) All that suggests that the reported metrics in most NRSfM papers reflect an upper-bound accuracy that cannot be reached in practice.

Second, NRSfM assumptions are often not fulfilled in practice, which results in corrupted shapes even on accurate point tracks. Moreover, most (if not all) sequences demonstrated in papers on dense NRSfM can be accurately initialised under the rigidity assumption [TK92]; otherwise, dense NRSfM would not perform well on them. Moreover, due to the severe ill-posedness of NRSfM, there is often no unique set of parameters (of the energy terms) working equally well across multiple datasets. Consequently, some recent research addresses scalability [AGS17, KCDL18].

Noticeable is also the saturation of the field of dense NRSfM. One of the main reasons is that the numbers are improving marginally on the existing datasets, let alone that such improvements can barely be noticed qualitatively. Most datasets contain small motions and are widely considered not challenging enough to boost the progress in dense NRSfM. Next, the notion *NRSfM* is being used in other contexts than originally meant. Consider so-called "deep NRSfM" methods for sparse 3D reconstruction from single images [NRG*19, PLK20, WL21, ZDY*21, SPJG22]. The underlying neural networks are trained on large image collections without 3D supervision and do not always use observed object motions as one of the 3D reconstruction cues. Moreover, these 2D-to-3D lifting methods often require different datasets for each object class.

One unsolved problem in the field remains dense NRSfM with shape completion. Since a single keyframe is selected for point tracking, only the points visible in it are subsequently reconstructed; the points that become visible in other frames are discarded. A naïve approach with shape completion would require multiple keyframes and a subsequent 3D surface fusion; no such technique has been demonstrated in the literature yet. Only recently first solutions to non-rigid shape estimation and completion from monocular videos were shown in the context of non-rigid neural radiance fields [TTG*21, LNSW21]. Thus, NR-NeRF [TTG*21] can simultaneously reconstruct a volumetric scene representation of a deformable scene from monocular videos (no 2D point tracks are required) so that all input views contribute to the canonical volume and complement the already available 3D densities. Ub4D [JHS*22] specifically targets explicit surface extraction and comes in the setting even closer to dense NRSfM. We next look at volumetric rendering methods that reconstruct non-rigidly deforming scenes using volumetric 3D representations.

3.1.3. Neural Rendering Methods

Neural radiance fields have introduced a new area of general dynamic reconstruction methods from a video that do not neatly fall into SfT or NRSfM. Crucially, these methods all *combine* naïve volumetric rendering and a neural scene parametrization, but differ widely in the specific kinds of input annotation used. Neither a template nor long-term correspondences are in principle needed as input and, unlike most prior work, they include the static background in the reconstruction, making these methods much more flexible and easier to apply in real-world settings. In contrast to most prior work, NeRF-based approaches tend to use density functions for geometry and not hard surfaces, allowing for some slack during optimization. While this slack enables almost photo-realistic novel-view synthesis, the underlying geometry is seldom evaluated as it exhibits, in most cases, rather low-quality mid-level and fine details. In addition, foggy artifacts may arise. Improving the quality of the geometry is thus central to move towards better reconstructions. For a detailed discussion, we refer to Tewari *et al.* [TTM*22].

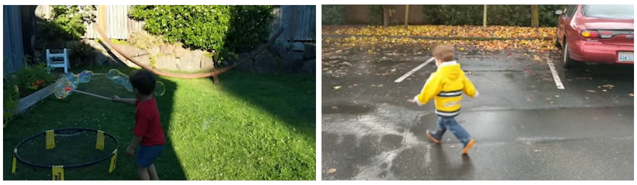

Figure 10: *Neural Scene Flow Fields [LNSW21] can model topology changes (bubbles) and view dependence (ground reflection). Image adapted from [LNSW21].*

State-of-the-Art Methods. Six concurrent works were the first to extend NeRF to the dynamic setting, covering a wide design space by choosing different trade-offs. They fall into two broad categories: time conditioning [LNSW21, XHKK21, DZY*21] and ray bending [PSB*21, TTG*21, PCPMMN21], which correspond roughly to the coordinate-system focus of Eulerian motion formulations and the particle focus of Lagrangian motion formulations in physics, respectively. The first category conditions the radiance field (a coordinate-based MLP parametrizing geometry and appearance) on a temporal input, *e.g.* time t, and thus loses long-term correspondences, which gives it the freedom to reconstruct large motions and topology changes. Consistency across time (via jointly optimized scene flow) is encouraged by warping losses, optical-flow losses, or keypoint losses. Thus, information is not propagated well in the long-term, restricting novel-view synthesis to nearby views at any time t. The second category disentangles the deformations into a separate, time-conditioned deformation field that acts on top of a static canonical radiance field, effectively bending rays to model deformations via space warping. Since this enforces hard correspondences across time (via the canonical model) and hence geometry and appearance information is shared across all time steps by design, it is empirically limited to a much smaller range of motion and does not cope well with topology changes but enables more challenging novel-view synthesis. Results from both categories exhibit close to photo-realistic appearance, although artifacts are noticeably more common than in the static setting, due to the more challenging nature of the problem.

Neural Scene Flow Fields (NSFF) [LNSW21] show that complicated real-world lighting effects like shadows and reflections in dynamic scenes can be modeled well by NeRF-like approaches, see Fig. 10. Nerfies [PSB*21] introduce an *SE(3)* deformation parametrization that is well-suited for deformations that are mostly articulated. Non-Rigid NeRF (NR-NeRF) [TTG*21] shows that a video captured by a moving camera with associated time stamps and camera parameters (and no other annotations) is sufficient to reconstruct scenes with small deformations. Xian *et al.* [XHKK21] show that recent depth-estimation methods [LHS*20] offer helpful guidance for reconstruction. NeRFlow [DZY*21] uses a Neural-ODE-based [CRBD18] deformation model, which is slow but invertible by construction and avoids self-intersections by design.

After this initial wave of works, progress has slowed recently, as noticeable improvements in this challenging setting, beyond mere shifting of trade-offs, have been hard to come by. HyperNeRF [PSH*21] is a follow-up to Nerfies [PSB*21] with a sophisticated conditioning of the canonical model, which is not only temporally but also spatially varying. This enables the reconstruction of topology changes and larger deformations than Nerfies but comes at the cost of losing correspondences. It is a hybrid of both categories. Gao *et al.* [GSKH21] introduce a new time-conditioned method that exploits single-view depth in a scale-invariant depth-order loss. Unbiased4D [JHS*22] steers NR-NeRF towards surface estimation by replacing the commonly used density function for geometry by an SDF, following NeuS [WLL*21]. Marching Cubes then allows to easily extract high-quality meshes from the reconstruction, although temporal correspondences are lost in that process. Fang *et al.* [FYW*22] apply a fast, explicit, voxel-based data structure to reduce training time from many hours to a few minutes. Guo *et al.* [GCD*22] speed up training similarly. They explicitly handle occlusions. Qiao *et al.* [QGL22] use differentiable mesh-based physics simulation as a soft constraint on the deformation field. Subsequently, they can edit the reconstruction in a physical manner.

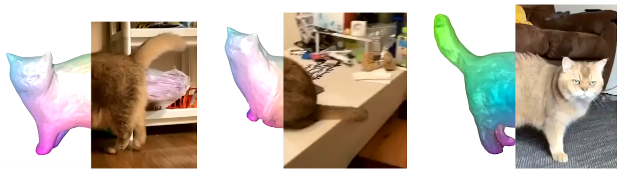

Figure 11: *BANMo [YVN*22] uses a NeRF-style object representation to reconstruct an object from a few monocular videos. Image adapted from [YVN*22].*

Datasets. So far, no standard datasets or benchmarks are established and all works evaluate predominantly on self-recorded scenes or, in some cases, the dataset from Yoon *et al.* [YKG*20]. The most recent work, Fang *et al.* [FYW*22], evaluates on synthetic scenes from D-NeRF and real scenes from HyperNeRF. Gao *et al.* [GLT*22] thoroughly analyze the limitations of currently used datasets.

3.1.4. Other Methods for Few-Scenes Reconstruction

There are a number of other reconstruction works that focus on a single or a few scenes but do not fall into any of the previously discussed categories. We group them together here since a per-scene parametrization (*i.e.* auto-decoding [PFS*19]) is still feasible in this problem setting. All in all, this is a nascent niche with a lot of unexplored potential. However, it can merge with the NeRF-style works (Sec. 3.1.3) for the foreseeable future, as BANMo [YVN*22] indicates, and ignore non-neural alternatives that could, for example, build on differentiable mesh rendering.

In their pioneering work [YKG*20], Yoon *et al.* primarily work with estimated depth maps to 3D-reconstruct a temporal sequence. A neural network fuses these depth maps consistently and with small scene flow into a novel view, filling in holes. Subsequent image warping of the RGB input images followed by a neural blending network enables novel view synthesis, which, however, does not give correspondences across time. The individual networks of the method need to be pretrained on synthetic or larger datasets. A few works in Sec. 3.1.3 evaluate on Yoon *et al.*'s dataset.

Yang *et al.*'s LASR [YSJ*21a] uses foreground masks and optical flow to reconstruct a general dynamic foreground object as a deforming mesh (initialized to a sphere). They are the first to exploit differentiable mesh rendering in the general dynamic per-scene reconstruction setting. The follow-up ViSER [YSJ*21b] additionally attaches features to the geometry and matches them to image features (the features are learned from scratch), which provides more robustness than LASR against appearance changes. Unlike LASR, ViSER can also reconstruct multiple videos of the same object at once. In a further follow-up, BANMo [YVN*22], they merge this line of work with NeRF-style volumetric rendering and neural parametrization. It heavily relies on matching pretrained image features to establish correspondences. Its results on a wide variety of general objects show decent mid-level details, with slight temporal jitter and unnaturally smooth deformations, see Fig. 11.

LASSIE [YHL*22] reconstructs animals from a small-scale image collection (∼30 images of different individuals of the same species) and hence does not exploit temporal information. It goes even further than BANMo and completely forgoes any appearance loss, relying entirely on pretrained features. Although the results are far from photo-realistic, they are promising given the very challenging setting that does not need input annotations of any kind.

3.1.5. Methods Using Data-Driven Priors

When working with large-scale image collections instead of a few scenes at most, per-scene parametrization ceases to be practicable. Methods in this category thus need to rely on a data-driven prior, see Sec. 2.3.5. While this setting has long been common for category-specific methods, it only recently turned out to be a viable path for general methods as well. The main trend in this area is a preference towards reducing the need for involved annotations and exploring alternative annotation settings like video rather than improving the quality noticeably beyond the quality of the initial work (CMR [KTEM18]). Another testament to the difficulty of the problem setting is that most methods stick with the CUB dataset [WBW*11] of different species of birds, which only have challenging deformations in their wings, which are barely reconstructed by any existing method. For an excellent table summarizing methods in this section, we refer to Table 1 in a recent survey [MJK*22] and to Table 1 in DOVE [WJRV21].

State-of-the-Art Methods. After the earlier work by Tulsiani *et al.* [TKCM16] reconstructing rigid categories by deforming a template, interest in general image-collection approaches has started growing with Kanazawa *et al.*'s [KTEM18] CMR method, which mostly shows results on CUB, see Fig. 12. CMR uses foreground masks and manually labeled semantic keypoints as annotations, and they regress camera pose, per-vertex offsets of a mean shape, and appearance. Their analysis-by-synthesis method uses differentiable rendering of the mesh. To handle occlusions, they exploit the left-right symmetry of birds and only predict one side. For appearance, they regress, for every pixel of a UV map, where that pixel should sample the input image to copy its RGB color from, the so-called *texture flow*, a technique that remains in wide-spread use in this line of work. This leads to a good appearance quality, while the predicted geometries are of rather coarse quality. Fine structures like legs or large deviations from the mean shape like open wings hardly exist. In a follow-up, Goel *et al.* [GKM20] (U-CMR) get rid off the need for keypoints. Since CMR uses SfM on the keypoints to obtain camera poses, U-CMR no longer has easy access to rough poses and they instead optimize for a per-image set of potential cameras, in auto-decoder fashion. The cameras thus estimated are of slightly lower quality than CMR's and, accordingly, the result quality remains, at best, comparable to CMR.

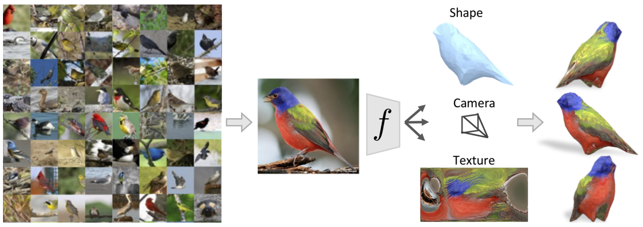

Figure 12: *CMR [KTEM18] learns a data-driven prior from a collection of images to reconstruct an object from a single image. Image adapted from [KTEM18].*

Building on an idea originally introduced by DensePose [RAG18], a couple of works start from predictions of (visible) object coordinates in image space, a dense analogue to sparse semantic 2D keypoints: In a follow-up to their work on rigid objects, Canonical Surface Mappings (CSM) [KGT19], Kulkarni *et al.* [KGFT20] (A-CSM) fit an articulated 3D geometry to object coordinates predicted in image space, with the articulations and coordinate predictions jointly learned without direct supervision on either. They require a template shape per category and can thus handle a wider variety of datasets than just CUB. While this makes their geometry inherently detailed, its deformations to fit to the input are rather coarse, often ignoring even legs in the input. Similarly, DensePose3D [SNG*21] exploits a pretrained DensePose model for humans and pretrained Continuous Surface Embeddings (CSE) [NNS*20] for animals to fit a skinned template to 2D object coordinates in image space. A canonicalization loss handles missing camera poses. Their result quality is similar to A-CSM, with only coarse deformations being somewhat accurate.

Tulsiani *et al.*'s IMR [TKG20] applies CSM to CMR's setting. Although still unpublished, it is widely considered as a proper member of this line of work. They allow for instance-specific offsets of the template before applying the skinning, which A-CSM does not. Their results contain legs and coarse deformations for a wide variety of animal species, but still severely lack in detail. Li *et al.*'s UMR [LLK*20] also no longer needs keypoints or camera poses, or even any kind of template. They obtain the same benefits that keypoints provide by exploiting self-supervised part segmentation in image space from prior work. While simplifying the required annotations, this yields quality on par with CMR. In their follow-up VMR [LLDM*20], Li *et al.* apply a standard per-image model to a video at test time. In addition, they no longer assume symmetry, instead replacing the single template with a linear subspace model obtained from clustering CMR's reconstructions. They exploit appearance constancy and the consistency of semantic parts to obtain a temporally consistent reconstruction. While their method makes the reconstructions less noisy and enables asymmetric deformations, the geometry and deformations remain coarse.

Wu *et al.* [WRV20] mainly exploit the symmetry of certain object categories like faces of humans and cats to reconstruct a canonical mesh, which is then rendered into an estimated camera view. Thanks to symmetry, a 2.5D mesh, and restricted input view points, they avoid having to meaningfully handle occlusions, and they hence do not need any kind of input annotation or template. Their results already exhibit decent mid-level detail, although the image resolution is rather low. DOVE [WJRV21], proposed by almost the same authors as the previous work [WRV20], is the first to use many videos at training time. Their goal is a per-image predictor at test time, the opposite of VMR's setting. This allows them to exploit temporal information via geometry and appearance consistency, and optical flow. Since they do not assume camera poses to be given, they argue that, for symmetric shapes, a simple flipping operation akin to their prior work [WRV20] is enough to decide between ambiguous poses instead of optimizing for a set of different cameras. Their results are of coarser quality than CMR's, since their input requirements are more relaxed.

Kokkinos *et al.* [KK21a] use a sophisticated deformation model of a template that is based on Laplacian deformations [SCOL*04] in an end-to-end differentiable manner. As they train their per-image predictor on videos, they encourage consistency with the optical flow between neighboring pairs of frames. Despite using keypoints, they find the camera optimization of U-CMR helpful. Crucially, at test time, they refine the predictions made by the neural predictor using auto-decoding-style instance optimization, similar to works in Sec. 3.1.4. On CUB, they can handle open wings but otherwise only barely improve the coarse geometry beyond the quality of CMR. In the follow-up TTP [KK21b], they turn around the correspondence regression of A-CSM and IMR, instead regressing the 2D UV coordinate for every vertex of the mesh. TTP trains a shared network for the UV regression task but performs end-to-end differentiable, iterative instance optimization to determine the deformation and camera parameters *at training time*. This noticeably improves their result quality over prior work, with the coarse geometry mostly correct and hints of mid-level details.

TARS [DP22] is the first work in this section to use a neural SDF parametrization rather than a mesh with fixed topology for the geometry. Similar to HyperNeRF [PSH*21] (see Sec. 3.1.3), TARS handles topology changes by conditioning the canonical model on a latent code. This lack of a shared canonical model (unlike prior work) improves the geometry quality on CUB, where open wings are now possible and some mid-level details are discernible.

3.2. Humans

Capturing the deforming 3D surface of humans from a single RGB camera, also called *monocular (human) performance capture*, has become a very active research area over the last decade. It complements and refines concepts initially introduced for the general case (Sec. 3.1). A key difference is that those methods introduce human-specific priors because the rough shape and topology remain the same irrespective of gender, age, and clothing type.

We categorize existing methods based on how strong their assumptions about the 3D geometry of the person are. Template-free methods do not assume prior knowledge of the specific 3D geometry (Sec. 3.2.2). Parametric methods leverage a low-dimensional parametric model of humans obtained by a database of 3D scans of thousands of humans (Sec. 3.2.3). Finally, template-based methods assume a pre-scanned 3D template of the person is given (Sec. 3.2.4). Before we review all categories in more detail, we introduce the problem-specific challenges in Sec. 3.2.1.

3.2.1. Challenges

On top of the general challenges (Sec. 2.6) of this inherently ambiguous setting, there are also human-specific ones. Humans are composed of individual body parts, *e.g.* arms and legs, which can move in a highly articulated and fast manner, leading to large displacements. This makes finding correspondences between neighboring frames non-trivial, and photometric consistency between a model and the input image can be challenging due to the local nature of image gradients. Moreover, the articulated structure can lead to severe self-occlusions, *i.e.* one body part is occluding another, and a sudden change in visibility can occur. This change in visibility is not only hard to track but also non-differentiable, and occluded body parts can only be tracked using priors. Last, there are two types of deformations for humans: The piece-wise rigid deformation induced by the skeletal pose and the non-rigid deformation of the surface, *e.g.* of the clothing. Both of them require special care and, at the same time, can only be considered jointly.

3.2.2. Template-Free Methods

Template-free methods do not assume the availability of a known template or a parametric model. Initial methods like BodyNet [VCR*18] and DeepHuman [ZYW*19] learned to reconstruct the human at the voxel level. However, such a representation is memory-intensive and suffers from quantization issues. To mitigate the memory issue, Moulding Humans [GFM*19] reconstructs humans by estimating the front and back depth maps. In a similar fashion, PeeledHuman [JCSN20] represents a human shape as a set of depth maps at the points of intersection of the camera rays with the human surface. While such representations use less memory compared to voxel-based reconstruction, they cannot account for high-frequency details due to the finite resolution of depth maps.

These issues motivated work towards learning implicit models to represent a human. Thus, PIFu [SHN*19] and PIFuHD [SSSJ20] learn a zero-level set of the surface that can represent high-frequency details, as the model learns a continuous representation. This allows for improved handling of hair and clothing deformations (see Fig. 1, bottom right). However, generalization to arbitrary poses (Fig. 13) remains challenging because the only global context provided to the network comes from the image features at the query point. To mitigate this, Geo-PIFU [HCJS20] adds a 3D U-Net branch that provides geometric features for a query point.

Several NeRF-based methods were recently proposed to reconstruct humans in 4D from a monocular video. Human-NeRF [WCS*22] learns appearance as a continuous field in a canonical space and learns a mapping from the motion field to canonical space using two modules. The first module learns the skeleton-level deformations and the second one accounts for non-rigid deformations by learning corrective offsets on top of those deformations.

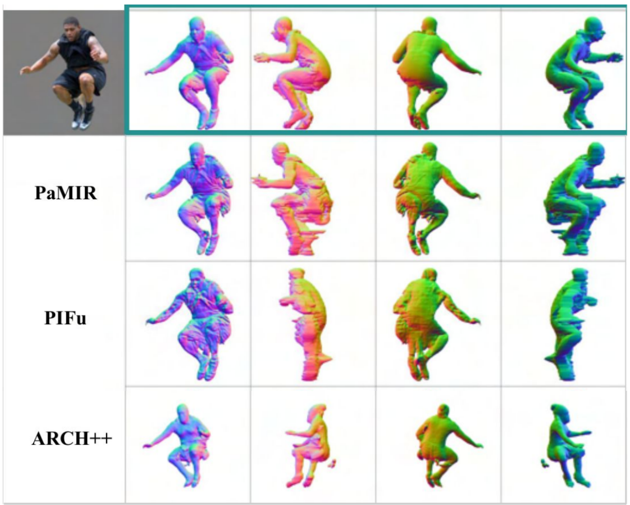

Figure 13: *Comparison of several monocular human reconstruction methods as illustrated in [XYTB22]. The top row corresponds to ICON [XYTB22]. Template-free methods like PIFu tend to struggle when tested on challenging poses. Image adapted from [XYTB22].*

Also related is PHORHUM [AZS22] that learns an SDF for the human body and based on pixel-aligned features, similar to PIFu. The method also estimates albedo with the same network and shading is estimated by a separate network using illumination features and the surface normals.

3.2.3. Approaches Using Parametric Models

Several methods use parametric models SMPL, SMPL-X, GHUM(L), or imGHUM [LMR*15, PCG*19, XBZ*20, AXS21] to estimate coarse pose-dependent geometry. They provide a topographically consistent canonical space and skinning weights.

Methods like MonoClothCap [XPWH20] optimize for per-vertex offsets from the SMPL template mesh. However, the optimization-based pipeline requires up to five minutes to reconstruct one frame. Methods like Tex2Shape [APMTM19] and Alldieck *et al.* [AMB*19] instead learn the per-vertex deformations and normals either directly from a UV-unwarped texture map obtained from the estimated SMPL mesh or from part-wise segmentation images. They can thus learn geometry in an image-to-image translation fashion, significantly reducing the inference time.

However, estimating the per-vertex deformations of a parameterized mesh inherently limits the level of high-frequency details that can be retrieved. This motivated several methods that learn an implicit representation based on a parametric mesh. Such methods first learn to map a point in the observation space to the canonical space of the parameterized mesh. Thus, piecewise rigid deformations are modeled using the skinning weights of the parametric model, and the non-rigid deformations are typically learned using a separate network. ARCH [HXL*20] proposes to learn the surface deformations as an implicit surface based on image features and a semantic deformation field that warps a posed mesh to the canonical space. ARCH++ [HXS*21] improves this by sampling a point cloud from the corresponding parametric mesh in a canonical space and then extracting spatial features using a PointNet++ Encoder. These spatially-aligned features, along with the pixel-aligned features from a UNet, are then fed to an occupancy network to learn the occupancy field. However, noisy observations can make it difficult to estimate the warping function. Alternatively, one can voxelize the estimated parametric mesh and extract the 3D voxel-aligned features from a 3D network, as done by PaMIR [ZYLD21]. Similar to the geometry-aligned features of Geo-Pifu, PaMIR proposes to use these 3D features in conjunction with the image features to learn an implicit 3D surface. ICON [XYTB22] learns an implicit 3D surface as a function of the front and back surface normal features and the SDF of the corresponding SMPL mesh. Recently, HF-Avatar [ZZL*22] proposes to produce high-fidelity by learning a reference-based neural rendering network and using it to refine the neural texture of the human in a coarse-to-fine manner.

Some methods split human performance capture into human body reconstruction and clothing reconstruction [BTTPM19, YPA*18]. MulayCap [SWY*22] estimates the parameters of the garments and uses them to re-dress the naked-body SMPL mesh using a simulator. The textures are reconstructed by rendering using the regressed albedo maps and shading images. There are also attempts towards learning a parameterized model of human clothing like SMPLicit [CPA*21]. However, parametric clothed human reconstruction is often restricted to a few clothing types and, generally, cannot span the space of all clothing items.

NeuMan [JYS*22] reconstructs the scene as well as the human by training a separate NeRF for each and jointly integrating samples from each in a ray. For human-specific deformations, NeuMan transforms the points in the observation space to a canonical space of an SMPL mesh using pose-dependent transforms. There are also multi-view neural-rendering-based methods like H-Nerf [XAS21] and Neural Body [PZX*21] that use the latent codes of imGHUM and SMPL, respectively, to train a NeRF (or SDF). However, their reconstruction quality significantly degrades when tested with a monocular input, even with relatively simple articulation.

3.2.4. Template-Based Methods

Finally, template-based methods assume a textured 3D geometry of the subject to be given. In contrast to the general case (see Sec. 3.1.1), *template* here refers to a textured 3D mesh of a clothed human. Typically, such a template is acquired by moving around the subject standing in a static T-pose and recording a monocular RGB video [XCZ*18, HXZ*19]. In an additional semi-automated step, the character mesh is then rigged and skinned to a kinematic skeleton. Given such a template and the RGB video of the person in motion, the goal of these works is to estimate the space-time coherent, dense, and non-rigid deformation of the 3D template.

The pioneering work MonoPerfCap [XCZ*18] is the first method that jointly tracks the skeletal pose and the dense surface deformation of the template from a single RGB video. In the first stage, they estimate the skeletal pose by regressing 2D and 3D joint predictions. Then, they fit the skeletal motion represented by a discrete cosine transform to the predictions by optimizing a non-linear energy function. The obtained skeleton motion is then used

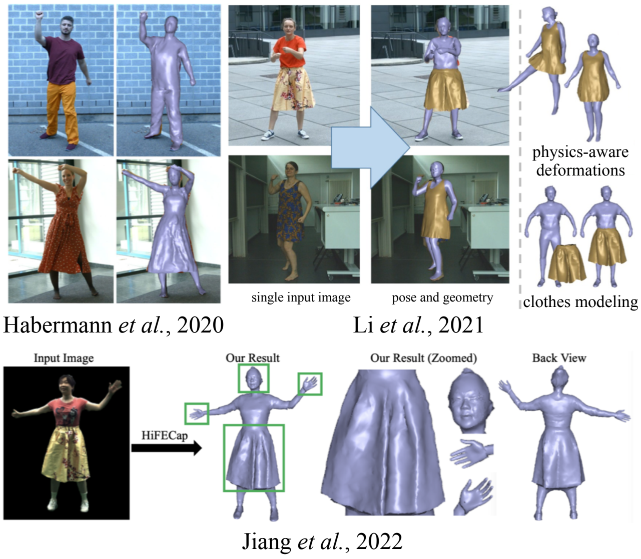

Figure 14: *Evolution of template-based human performance capture methods. Images adapted from [HXZ*20, LHT*21, JHGT22].*

to coarse-deform the template geometry and, in a second stage, further refine the surface deformations by fitting the geometry to the human silhouette. Although this work can achieve robust and temporally smooth results, the overall method requires more than a minute to optimize a single frame. LiveCap [HXZ*19] is a more efficient approach for optimizing motion and surface deformation. It introduces an updated energy formulation and skeleton motion representation in conjunction with dedicated GPU solvers and a multi-threaded CPU pipeline to become the first real-time method.

Nonetheless, it remains challenging to achieve high 3D accuracy due to the depth ambiguity, and occluded surface parts are mostly driven by geometric priors and not any data terms. To overcome this limitation, DeepCap [HXZ*20] proposes to train skeletal pose and surface deformation networks, which take as input a single RGB image. During training, these networks are weakly supervised on multi-view images, which allows supervising surface areas that are occluded in the input view and also improves 3D accuracy.

The methods discussed so far all treat the template as a single connected surface, which does not reflect reality since clothing and the human body are separate. Thus, shifts of clothing along the body cannot be tracked well, and the cloth deformations usually contain baked-in wrinkles from the static template and do not look physically plausible as can be seen in Fig. 14. Li *et al.* [LHT*21] propose to separate the geometry into two layers, *i.e.* clothing and the human body. While the pose and deformation networks are leveraged from DeepCap [HXZ*20], they introduce a physics simulation layer, which enforces more physically plausible deformations during training and prevents cloth-body surface penetrations.

HiFECap [JHGT22] is the first method that jointly tracks the deforming clothing, the body pose, hand gestures, and facial expressions. They introduce a hybrid neural network architecture consisting of image and graph convolutions to better recover surface details. They demonstrate how existing parametric hand and face models can be fit onto a 3D template, and how those can be jointly deformed with the surface deformation of the clothing.

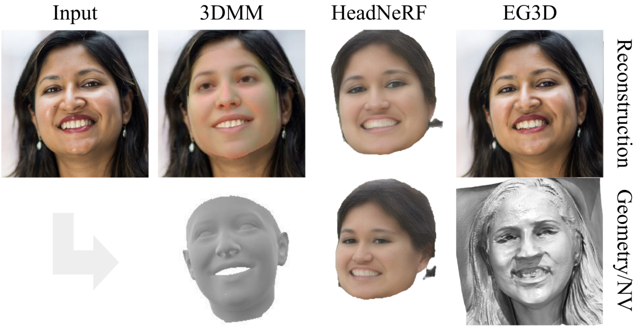

Figure 15: *Results of representative methods using an explicit 3DMM [LMFV*23], trained on multi-view data (HeadNeRF [HPX*22]) or in adversarial manner on large-scale monocular dataset (EG3D [CLC*22]). We show (top) the reconstruction and (bottom) its geometry or novel view synthesis (for HeadNeRF). Image adapted from [CLC*22].*

3.2.5. Future Directions

While parametric models of clothing *geometry* have been studied recently, creating a parametric geometry and appearance model of the whole human body remains an open challenge. This is due to a large amount of data necessary to sufficiently sample the model space. However, recent progress in dataset [HYH*20, CRZ*22] acquisition may now enable the building of such a model. Another unsolved fundamental problem is the tracking of topological changes (*e.g.* the person is taking off their jacket), while maintaining correspondences over time. Recently, implicit human models have been extensively researched, which can deal with topological changes due to their implicit representation. However, they lack space-time coherent correspondences. Complementary, explicit mesh models have also been studied. While maintaining correspondence naturally, they fail to faithfully track topological changes or surface details. Thus, in the future, a combined representation could lead to the best of both worlds. Moreover, the joint capture of all aspects of the human is still in its infancy, *i.e.* tracking of hands, face, body pose, clothing, hair and eye gaze. While the solution for individual body parts exists, it remains an open question of how they can be efficiently and effectively combined for real-time performance. The detailed tracking of hair is another open challenge since its thin and highly dynamic structure is not suited for surface-based methods. Thus, future research may involve alternative representations for hair that enable space-time coherent tracking. Finally, the robustness and interpretability of results are still a problem for learning-based approaches. Here, physics could improve the performance further, as seminal works already show [LHT*21, SGXT20, SOC22, YZH*22].

3.3. Faces

3D reconstruction of faces from monocular images is a heavily researched topic. In contrast to many other object types, faces have such advantageous properties as symmetry, small deformations and well-defined keypoints that can be exploited in the ill-posed 3D reconstruction setting. Facial shapes can be modeled in low-dimensional spaces and with linear models. Simple and effective models like PCA-based ones lead to reasonable results and are currently state-of-the-art in the monocular inverse rendering setting. Fueled by the availability of large amounts of data, this leads to faces being one of the dominant applications in the community.

This section provides an overview of recent developments, datasets, and applications of monocular 3D face reconstruction. Whilst some methods focus on photorealism and learned representations, other applications exploit a parametric representation and benefit from a classical statistical prior. The following parts are structured focusing on the distinction between classical explicit (Sec. 3.3.1) and modern implicit models (Sec. 3.3.2). We also cover specialized models for facial parts and the recent new dataset (Sec. 3.3.4). Moreover, Tab. 2 categorizes the covered methods.

3.3.1. Explicit Morphable Models

3D Morphable Models (3DMMs) [BV99] are statistical models of face shape and appearance variation with an explicit surface representation. They are built from a (comparably) small set of hundreds of faces and can be used as a prior for non-rigid 3D face reconstruction. There are recent surveys on 3DMMs [EST*20] and monocular 3D face reconstruction and tracking [ZTG*18], and we here focus on the recent developments arising after these surveys.

The core application area of explicit 3DMMs is the 3D reconstruction of faces from single 2D images through inverse rendering. While this problem has been studied extensively [EST*20], most evaluations were performed qualitatively and in highly constrained scenarios. One current trend in 3D face reconstruction with 3DMMs is their application to in-the-wild images. The NoW challenge [SBFB19] provides, for the first time, a way to quantitatively evaluate dense 3D reconstructions on in-the-wild images. The NoW challenge contains 2054 2D images of 100 subjects and ground-truth 3D scans of each person. Notably, the 3D scans were captured in a studio and not at the same moment as the images. Most methods participating in the NoW challenge are unsupervised or weakly supervised methods that do not exploit pairs of 2D images with 3D geometry. Whilst such data exists for controlled lab conditions, it is not publicly available for the in-the-wild setting.

The state-of-the-art methods on the NOW challenge and in the unsupervised setting are DECA [FFBB21] and FOCUS [LMFV*23], respectively, both 3DMM-based. Notably, no implicit modeling approach has participated in NOW so far. DECA exploits weak identity supervision and learns on videos and images. It goes beyond the simplistic linear 3DMM space and adds fine details to the 3D reconstructions through a displacement map trained with a detail-consistency loss (to separate person-specific details from a generic learned expression model). FOCUS, instead, learns without identity supervision and achieves similar performance by learning a robust model estimation and being robust to occlusions. The recent MICA approach [ZBT22] is trained on paired 3D and 2D data in a fully supervised way. Whilst the original NoW challenge ignores the scale of the reconstructed face, MICA can reconstruct the face shape well due to the 3D supervision; it outperforms unsupervised methods by a large margin under the metrical evaluation protocol.

TRUST [FBT*22] by Feng *et al.* (see Fig. 1, bottom row; second on the right) is the first method explicitly aiming at correct skin tone estimation based on weak supervision through multiple faces in an image and assuming a constant illumination condition. Their FAIR dataset addresses biases in 3D face reconstruction regarding skin tones and ethnicity [FBT*22]. It provides ground-truth albedos of synthetic faces to evaluate albedo reconstruction, focusing on disentangling diverse skin tones and lighting conditions.

Notably, explicit 3DMMs based on PCA [BV99] are still state-of-the-art for both the NoW and the FAIR benchmark, despite the extensive research done in the area of implicit models in recent years (Sec. 3.3.2). We assume this is due to the difficulty of inverting implicit models in challenging scenarios that include extreme poses, illumination, or occlusions and their comparably expensive rendering. Whilst model inversion was trivial for the Eigenfaces [SK87, TP91] approach, it became increasingly difficult for active appearance models [CET98] and is still a research topic for explicit and implicit 3DMMs.

3.3.2. Implicit Morphable Models

Recently, there has been a flurry of methods based on implicit representations. As compared to mesh-based representations, implicit models are not restricted to a fixed topology and, as a result, can model the entire head, including hair. As these methods also model the face appearance using neural networks (which generally have a much better capacity than simple linear models of 3DMMs), they can synthesize photorealistic faces. It is observed that most of the methods based on implicit representations target applications that require photorealistic renderings, with very few methods targeting accurate geometry estimation [RTE*21, ZYHC22].

We can broadly categorize implicit models based on the type of training data. Some methods use posed multi-view image sets of several identities [RTE*21, HPX*22, WCZ*22, ZZSC22], the others use monocular images of several identities without paired camera poses [WRV20, CMK*21, CLC*22, DYXT22, XYDT22, GLWT22, OELS*22], and several approaches work with video data [ZAB*22, GPL*22, GTZN21, AXS*22]. Methods that use data of multiple identities at training time usually build a prior of face shape and appearance with implicit representation using a neural network to parameterize the face space. These models act as strong priors at test time to reconstruct any face, also from a single image. Person-specific methods require video data of a single person's face and can reconstruct the entire scene through time. Very few implicit-based methods address the correspondence problem [WCZ*22, TBP*22]; correspondences are necessary for downstream applications such as texture transfer or registration.

Multi-View Supervision. Several approaches use multi-view data of multiple identities to learn face priors, mostly using neural networks. These methods draw inspiration from the auto-decoder architecture of DeepSDF [PFS*19]. Instead of a separate neural network representing each sample, DeepSDF showed that it is possible to learn the entire space of an object category by conditioning the network output with a latent vector specific to each object sample.

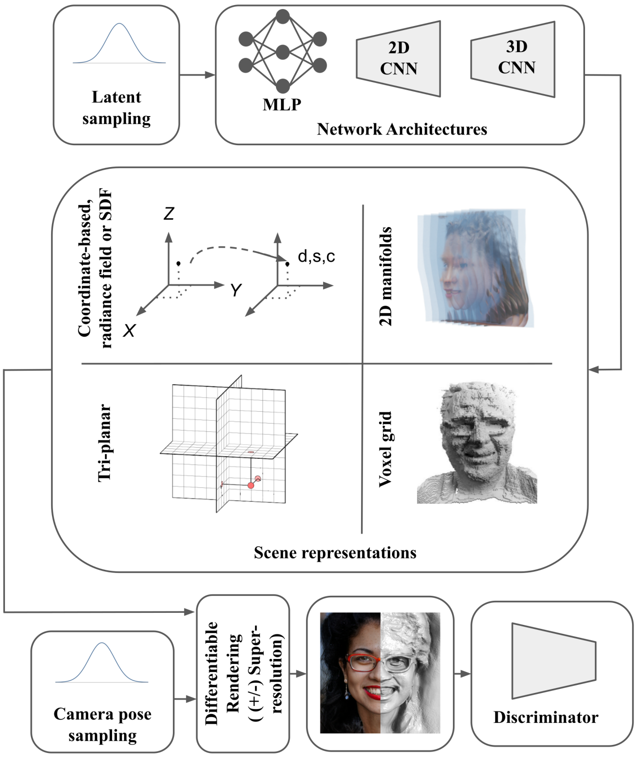

Figure 16: *A typical pipeline for 3D-aware GANs. The generator can be parameterized by an MLP [SLNG20, CMK*21], a 2D CNN [CLC*22] or a 3D CNN [XPY*22], to regress a radiance field [SLNG20, CMK*21], SDF [OELS*22], 2D manifolds [XYDT22], tri-planar [CLC*22] or voxel-based scene representation [XPY*22], which is trained with an image discriminator. Images adapted from [XYDT22, CLC*22, SSN*22].*

During training, the latent vectors are also learned along with the network parameters. Then, at test time, an unseen sample is reconstructed by optimizing for the latent vector.

HeadNeRF [HPX*22] extends this strategy to face data and builds autodecoder models using data from multiple identities. As their training dataset also contains each identity in multiple expressions, they can disentangle deformations due to facial expressions from identity-specific deformations by having a separate latent vector for expressions. MoFANeRF [ZZSC22] also has similar design choices with one exception, *i.e.* the training is performed in a feed-forward manner with identity and expression parameters estimated with the help of 3DMM. Since this model is trained on data without hair and relies on estimated 3DMM parameters only, they cannot model hair. MoRF [WCZ*22] extends a similar approach to include more features: They learn to map each training identity into a canonical space, use registered meshes for guidance and model diffuse and specular components explicitly.

2D Supervision. As obtaining large-scale multi-view data is challenging, most methods mentioned above are trained with <410 identities [HPX*22], which impacts generalizability. We next discuss methods for building face priors from large-scale monocular data. Obtaining monocular data is easier than collecting multi-view data, these methods are often trained with more than $7 \cdot 10^4$ identities [EST*20] and, as a result, can generalize better.

Some methods do not assume the camera poses to be given as input [SLNG20, CMK*21], while others require them [RMY*22]. The latter methods learn the face model in an adversarial manner (Sec. 2.5) and often use a generative scene model with a 3D representation (*e.g.* radiance fields or SDF) parameterized by a latent space. During training, they assume a known distribution for camera poses and a fixed latent space—that they sample in every iteration—and render the scene to synthesize 2D images. The models are then trained with the help of a discriminator. Please refer to Fig. 16 for a typical 3D-aware GAN pipeline.

GRAF [SLNG20] and pi-GAN [CMK*21] are the first methods to build a generative model with NeRF [MST*20] in an adversarial manner. The sampled images are not as high-quality as image-based generative models [KLA19] because of deficient sampling in the volumetric integration of rays. Moreover, the Monte-Carlo-based sampling results in ineffective training [DYXT22]. GRAM [DYXT22] overcomes this limitation by learning the radiance fields only on a set of 2D manifolds—which are common across different identities— improving the quality of rendered images. However, the learned manifolds are biased toward frontal images, as the dataset primarily consists of frontal-looking images. Extreme novel views contain severe artifacts. LOLNeRF [RMY*22] shows that learning 3D head models from large-scale monocular image collections is also possible using image reconstruction loss instead of purely adversarial loss. It expects paired camera pose as input, which they obtain from predicted keypoints. However, the random samples from the learned model are not as photorealistic as the ones trained in an adversarial setting.

The methods discussed above need to query coordinate-based MLPs for many points on all the rays to render the full image; they can train models with up to 256×256 resolution. Other recent methods [OELS*22,GLWT22,XYDT22,XPY*22] try to overcome this limitation, *i.e.* they maintain 3D representation at a lower resolution and apply a super-resolution module that takes the rendered 3D data to synthesize high-resolution 2D images. Still, this policy is not truly multi-view consistent as the super-resolution module operates in 2D. To overcome the limitation of not being able to train implicit models at high-resolution because of computational complexity, EpiGRAF [STWW22] proposed a novel space and scale aware discriminator which enable patch-based training of the generator model. Recently proposed D3D [TBP*22] learns a canonical space of faces without supervision, which helps in downstream tasks like color and segmentation transfer between faces. GAN2X [PTLT22] utilizes StyleGAN2 [KLA*20] to create pseudo-multi-view images for a given input image and has an explicit 3D-to-2D image formation model. The obtained labels are used to learn geometry, appearance, and illumination parameters in an iterative manner. The discussed methods model face as a single entity, although it has multiple semantic parts. gCoRF [BTP*22] addresses this concern by explicitly representing each part of the face by a separate 3D representation. This enables exciting applications, such as editing facial regions in volumetric space.

Monocular Video. Several methods require a monocular video of a

Object Type	Representation	Prior	Training Data	Other Features
only face: [FFBB21, LMFV*23] full head: [HPX*22, ZZSC22, WCZ*22, RTE*21, SLNG20, CMK*21, DYXT22, RMY*22, OELS*22, GLWT22, XYDT22, XPY*22, TBP*22, PTLT22, BTP*22] hair: [CSW*16, LHM*18, HSW*17, ZHX*18, SHM*18, YSZZ19] eyes: [BBGB16, WBM*16, PVO*20] ear: [ZEJ*16, DPS18, PVO*20]	implicit: [HPX*22, ZZSC22, WCZ*22, RTE*21, SLNG20, CMK*21, DYXT22, RMY*22, OELS*22, GLWT22, XYDT22, XPY*22, TBP*22, PTLT22, BTP*22] explicit: [FFBB21, LMFV*23, CSW*16, LHM*18, HSW*17, ZHX*18, SHM*18, YSZZ19, BBGB16, WBM*16, PVO*20]	adversarial: [SLNG20, CMK*21, DYXT22, RMY*22, OELS*22, GLWT22, XYDT22, XPY*22, TBP*22, PTLT22, BTP*22] PCA: [FFBB21, LMFV*23, BBGB16, WBM*16, PVO*20, ZEJ*16, DPS18, PVO*20] exemplar: [CSW*16, LHM*18, HSW*17]	multi-view: [HPX*22, ZZSC22, WCZ*22] monocular images: [SLNG20, CMK*21, DYXT22, RMY*22, OELS*22, GLWT22, XYDT22, XPY*22, TBP*22, PTLT22, BTP*22] monocular video: [GTZN21, AXS*22, ZAB*22, GPL*22]	canonical space: [WCZ*22, TBP*22, FFBB21, LMFV*23] compositionality: [PVO*20, BTP*22] reflectance properties: [WCZ*22]

Table 2: *Overview and classification of face-specific methods.*

Dataset	Format and Resolution	Coverage	Samples	Scanner
FaceScape [YZW*20]	triangle mesh (2M vertices), texture images (resolution 4096 × 4096), raw camera images (359 id × 20 ex × 60 views in 4M-12M pixels)	full head including face, neck, ears, excluding eyes	938 individuals × 20 expressions	multi-view system with 68 cameras
Multiface [WZA*22]	triangle mesh (7306 vertices), texture images (resolution 1024×1024), raw camera images (2048×1334), including audio	full head including face, neck, ears	13 individuals × 65 (v1), 118 (v2) expressions	multi-view system with 40 (v1) to 160 (v2) cameras
H3DS [RTE*21]	triangle mesh (120k vertices), texture images (resolution 2048 × 2048), raw camera images (512 × 512)	full head including face, neck, ears, eyes closed	23 individuals	structured light, multi-view (68 cameras)
CelebV-HQ [ZWZ*22]	monocular video dataset (512 × 512), with audio, manually annotated 83 facial attributes	full head including face, neck, ears, eyes	15653 individuals	monocular camera

Table 3: *Overview of publicly available human face datasets (extends Table 1 from Egger et al. [EST*20]).*

person during training to recover geometry and appearance. As they can learn from multiple video frames, they typically can capture high-quality face geometry and appearance. However, the quality comes at the cost of collecting a person-specific video.

NerFAC [GTZN21] models videos with dynamic faces with a 3DMM. 3DMM helps them bring the rays to a canonical space with a rigid transformation, and they learn a neural network in this space conditioned on tracked expression parameters to regress the radiance field for each frame. RigNeRF [AXS*22] uses a similar approach—but with an explicit deformation field as a function of expression parameters—to bring to the canonical space instead of naively conditioning the neural network as in NerFACE [GTZN21]. Recently proposed I M Avatar [ZAB*22] utilizes occupancy fields to model geometry. The critical contribution of this method is an analytical gradient formulation for the iteratively located surface intersection via implicit differentiation, which allows for end-to-end training. It also makes use of a tracked mesh using a 3DMM along with per frame delta blendshapes and skinning weights to bring the points to canonical space, in which the texture is modeled. The method of Grassal *et al.* [GPL*22] utilizes an explicit mesh-based model to address a similar problem. Along with a base geometry, which they get from a tracked face using a 3DMM, they also predict vertex offsets as a function of the head pose. This makes the method more compatible with the traditional graphics pipeline.

3.3.3. Specialized Models of Face Parts

Faces are complex, and some facial components that are hard to model with global models are targeted with specific models. The data availability is a key difference between such models in contrast to whole-face models. Whilst various datasets are available for faces, there are very few shared datasets for individual facial regions. This results in slower development of specialized models, and we observe that state of the art in the monocular setting is not yet using modern learning and neural rendering techniques.

The initial methods [CSW*16, LHM*18, HSW*17] for monocular hair reconstruction relied on a database retrieval. In contrast, recent methods train neural architectures to regress hair shape directly [ZHX*18, SHM*18, YSZZ19]. The approaches targeting high-quality eye and ear reconstruction follow face 3DMM methods by building separate 3DMMs for eyes [BBGB16, WBM*16, PVO*20] and ears [ZEJ*16, DPS18, PVO*20].

3.3.4. Data

A recent survey [EST*20] summarized the publicly shared face datasets at the time. We, therefore, focus on the datasets that arrived since (following the format of Table 1 in [EST*20]) and present our extension in Tab. 3. Along with 3D data (multi-view images), we also discuss a monocular video dataset [ZWZ*22], as there have been methods in the past taking advantage of video datasets for learning 3DMMs of faces [TBG*19, BTS*21].

3.3.5. Limitations and Outlook

Most methods that rely on explicit 3DMMs fail to capture fine-scale geometric details of faces that are perceptually important. Recently proposed implicit methods trained in an adversarial manner (3D-aware GAN) show promising results in obtaining some fine-scale details [CLC*22]. We show qualitative comparisons in

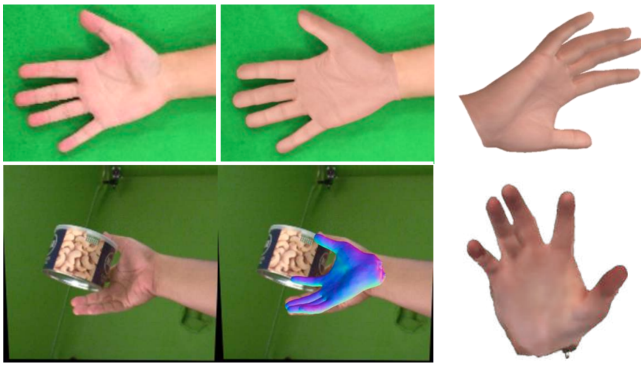

Figure 17: *Parametric hand models HTML [QWM*20] (top row) and LISA [CHV*22] (bottom row) support hand texture. Row-wise from left to right: The input image, the 3D hand reconstruction overlaid on the input, and a novel view of the reconstruction. Images adapted from [QWM*20, CHV*22].*

Fig. 15. Note that no existing monocular implicit methods quantitatively evaluate the 3D reconstruction accuracy w.r.t ground-truth 3D shapes. Moreover, since all the implicit methods are generally over-parameterized with neural networks, they can overfit to test images by baking in geometric details into texture space. They also can converge to inaccurate geometry if the initialization during test time is incorrect. It is also observed that methods with implicit representations, which take advantage of a large-scale video dataset [ZWZ*22], are under-explored compared to methods with explicit representation. This could be an interesting direction to achieve higher performance in capturing especially better facial expressions by exploiting the nature of the data.

3.4. Hands

Similar to human bodies, human hands are articulated objects with pose-dependent deformations on a fine scale. In contrast to human bodies, hands cause more severe self-occlusions and do so more often, especially in the monocular setting. Simple hand movements can be densely tracked in 3D by SfT methods [VA13, YRCA15]. However, this requires a known 3D template of the observed hand in advance. Moreover, SfT methods are not robust to large self-occlusions that are typical for hands. A stronger 3D shape prior can help to mitigate these challenging self-occlusions and appearance variations, *i.e.* a statistical parametric hand model covering the entire space of hand shapes [RTB17, QWM*20].

Single Hands. There are several approaches for 3D shape and pose estimation from monocular inputs [BBT19, BKK19, ZLM*19, GRL*19, ZCY*19, ZHX*20]. They all regress parameters of the MANO model [RTB17] and differ in their architectures, supervision, and fine-tuning policy for in-the-wild data. These methods rely on a differentiable mesh renderer [BKK19] or depth map rendering [GRL*19], differentiable re-projection loss [ZLM*19] or re-projection of 3D hand joints to images (2D keypoints) [BBT19, ZCY*19, ZHX*20]. All of them train on synthetic or mixed datasets with ground-truth 3D hand meshes and poses and some fine-tune on in-the-wild images using either 2D annotations only [BBT19] or rendered depth maps [GRL*19]. Moreover, MANO differentiability enables end-to-end trainable architectures [BBT19, ZLM*19]. Further characteristics of these methods are: Boukhayma *et al.* [BBT19] employ a convolutional encoder and a fixed MANO-based decoder; Baek *et al.*'s hand mesh estimator (HME) [BKK19] is supervised by 3D skeletons and hand silhouettes; Ge *et al.* [GRL*19] use a GCNN for mesh generation; Zhang *et al.* [ZLM*19] regress camera and mesh parameters with an iterative regression module; finally, Zhou *et al.* [ZHX*20] apply an inverse kinematics network, for the first time in the context of hands. Moreover, their proposed decoupling of image-to-keypoint and keypoint-to-angles regression allows them to train on all available data modalities, *i.e.* 2D and 3D annotated image data as well as pure motion capture data without paired images.

The methods discussed above estimate hand shapes from single images independently; their results on videos can be jittery. In contrast, the SeqHAND approach [YCLK20] integrates temporal consistency constraints by learning visual and temporal features from a synthetic dataset mimicking hand movements. Noteworthy is their synthetic-to-real fine-tuning policy involving detaching the recurrent layer from the core architecture and replacing the video input with single real images. A recent transformer-based work by Park *et al.* [POM*22] on 3D mesh estimation targets robustness against occlusions. They are interested in scenarios with hand-object interactions and treat objects as occluders.

Several works improve upon different aspects of the MANO model. HTML [QWM*20] is the first parametric hand texture model and it is learned from over one hundred SfM scans representing people of different genders, ages, and skin colors. HTML can regress both shapes and texture thanks to an analysis-by-synthesis photometric loss. Such a loss affects the shape estimates due to the additional supervision signal (the re-projected texture). LISA [CHV*22] is another hand model based on MANO that supports hand textures, but it uses different shape parameters and the shapes are represented by implicit functions. It has a disentangled parameter spaces for texture, shape, and poses learned from multi-view RGB videos annotated with 3D joints. Like HTML, LISA can reconstruct hands from monocular RGB images; see Fig. 17. DeepHandMesh [MSL20] is a neural encoder-decoder that leverages a personalized hand model (*i.e.* assuming the same subject at training and test) trained in a weakly-supervised manner from multi-view depth maps. It addresses MANO's limited resolution and implements a penetration avoidance loss to make hand-surface interactions more physical plausible. However, more identity-specific geometric details require a new training dataset for each identity and limit the generalizability to other identities.

Two Hands; Hands and Objects. Reconstructing two interacting hands adds complexity to the problem due to mutual hand occlusions and hand interactions that affect the surfaces of both hands. Only recently, the first solution to this challenging task has been introduced by Wang *et al.* [WMB*20] (Fig. 18-(left)). Their RGB2Hands method takes inherent depth ambiguities and mutual hands occlusions into account. It intermediately estimates segmentation for the handiness, inter-hand relative depth, and inter-hand distances. However, intertwined fingers can lead to hand-hand penetrations. The follow-up HandFlow [WLM*22] predicts a distri-

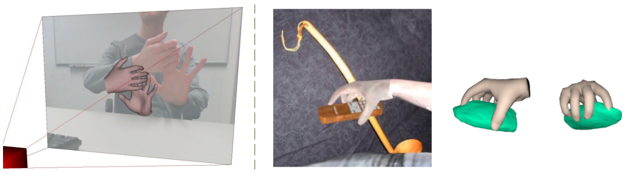

Figure 18: *Monocular 3D reconstruction of two interacting hands (RGB2Hands [WMB*20]) and joint hand-object 3D reconstruction (Grasping Field [KYZ*20]) are emerging directions. Images adapted from [WMB*20, KYZ*20].*

bution of plausible hand poses instead of a single estimate. The authors highlight that current evaluation schemes assuming a single correct hand pose are deficient. Zhang *et al.* [ZWD*21] leverage a hand-pose-aware attention block for per-hand feature extraction and a cascaded refinement block. The latter improves the initially estimated hand poses and shapes in the MANO space taking into account the interaction context between two hands. The proposed method achieves state-of-the-art accuracy and improvements in scenarios with inter-hand occlusions. Keypoint Transformer [HSRL22] predicts 3D poses of objects and hands observed in a single RGB image. The method includes three stages: It first detects and disambiguates the hand keypoints using a self-attention mechanism and then estimates the 3D hand poses with a cross-attention module. Another recent work [LAZ*22] further advances the two-hand case with the help of GCNNs and two attention blocks and shows a live demo of the proposed method. It accurately reconstructs challenging in-the-wild images with inter-hand occlusions. As of this writing, the last two discussed methods are the most accurate on the InterHand2.6M benchmark in the literature.

The joint tracking of hands and objects is an emerging area. Karunratanakul *et al.*'s Grasping Field [KYZ*20] is a new joint representation for hands, objects and the contact areas using implicit surfaces. They propose a neural method for hand-object reconstruction, assuming that a 3D model is given as input; see Fig. 18-(right). Hasson *et al.* [HVT*19] jointly reconstruct the shapes of a hand and an object after training on a new synthetic dataset. They argue that object manipulation simplifies the problem by providing more constraints and show that it improves grasp metrics. A follow-up work [HTB*20] assumes that a 3D model of the observed model is given. Ye *et al.* [YGT22] make a related observation that hand articulations are driven by local object shapes. Starting from the input image and hand and camera poses estimated by an off-the-shelf system [RSJ21], they reconstruct the object shape with an SDF decoder for the object shape. Like Hasson *et al.* [HVT*19], they encourage contact between the hand shape and the object at pre-defined regions [HVT*19].

Datasets. Only a few datasets in the literature provide RGB images and corresponding 3D shape annotations. FreiHAND is the benchmark for 3D hand pose and shape estimation of a single hand [ZCY*19]. ObMan [HVT*19] and DexYCB [CYX*21] contain shape annotations for single hands and objects. While ObMan provides single synthetic images, the more recent DexYCB includes videos of real grasping scenes recorded from multiple views. Moon *et al.* [MYW*20] introduce the InterHand2.6M dataset. Mesh annotations for it are also available thanks to NeuralAnnot [MCL22]. MultiHands [WLM*22] is an extension of InterHand2.6M with 100 additional annotations per image, which allows quantifying pose ambiguities as a distance between the predictions and ground-truth pose distributions. Finally, H2O is a popular dataset with shape annotations for two hands manipulating rigid objects [KTS*21], and H_2O-3D is currently the most challenging dataset with accurately annotated videos of two hands manipulating objects (due to large mutual occlusions caused by hands and objects).

Future Directions. Existing models lack geometric and pose-dependent texture details (*e.g.* nails, hair and blood vessels). We will soon see new methods for the 3D shape estimation of 1) hands and articulated objects and 2) hands and deformable objects. Moreover, reconstruction under various illumination conditions remains not solved satisfactorily, and reconstruction of interacting hands can advance further by improving mesh collision handling.

3.5. Animals

Unlike Sec. 3.1.4 and 3.1.5, this section discusses animal reconstruction methods that use parametric models. Apart from the seminal work by Cashman *et al.* [CF13] reconstructing dolphins, interest in animal-centered reconstruction has started growing only recently with the introduction of the SMAL model [ZKJB17], a SMPL-style model for quadrupeds learned from 3D scans of animal toys. It enables sufficient regularization to cope with the lack of large, high-quality datasets as are widely used in the mature areas of face and human reconstruction, which is both due to less *a priori* interest and the difficulty of capturing a wide variety of animals in a highly controlled setting. Biggs *et al.* [BRFC18] fit the SMAL model to videos instead of images. Zuffi *et al.*'s [ZKB18] SMALR uses keypoints and silhouettes to deform the SMAL model with per-vertex offsets beyond the parametric shape space. In the follow-up work 3D Safari [ZKBWB19], they train a regression network on synthetic Zebra images and apply it to real data without annotations at test time. Dogs have also received some attention: Biggs *et al.* [BBC*20] add limb scaling to create a dog-specific SMAL model from internet images annotated with 2D keypoints and silhouettes. Li *et al.* [LL21] use graph convolutions in a hierarchical manner to refine a regressed SMAL mesh with per-vertex deformations. Most recently, BARC [RZSB22] turns SMAL into a breed-aware dog model by exploiting breed labels in a triplet loss. Another line of work focuses on birds: Badger *et al.* [BWM*20] build a SMPL-style parametric model of cowbirds without access to 3D scans, which is then applied to monocular regression of its parameters. Wang *et al.* [WKDB21] generalize this model to multiple bird species (see Fig. 1, top row; second from the left). Data-driven general methods from Sec. 3.1.5, like CMR [KTEM18] in Fig. 12, often evaluate on the CUB birds dataset and do not use a parametric model, which leads to very coarse reconstructions. For more discussion on bird reconstruction, we refer to [MJK*22]. Wu *et al.* [WCL*22] generalize across species by first retrieving a rigged template mesh via CLIP features [RKH*21] from a template database.

4. Discussion and Open Challenges

We next elaborate on current challenges in the field and discuss two nascent but promising future directions: methods using event cameras and physics-aware approaches.

Large Scale. While static methods [ZRSK20] can handle large scenes, only a few recent dynamic methods cope with a static background [LNSW21, PSB*21, YKG*20, JYS*22]. Distant background, even if static, is hardly reconstructed by current methods.

Multiple Objects. Static methods already scale to scenes with multiple objects [OMT*21]. However, handling multiple dynamic objects in the same scene is still in its infancy [MLS*22, MSL*23].

Editability. Beyond mere reconstruction, the ability to edit the scene's deformations, geometry, and appearance would enable the easy creation of digital assets (*e.g.* for interactive AR/VR). Classical geometry and appearance representations already possess an extensive toolbox for editing. However, deformations remain challenging to manipulate, especially for non-expert end users. While driving coarse deformations by re-posing the skeleton of a skinned template is relatively straightforward, creating the corresponding finer deformations (*e.g.* of cloth), remains difficult. Scene editing becomes even more challenging when using volumetric representations, like modern neural parametrizations for geometry and appearance: The latter use backward deformation models, where manipulation is less intuitive and more involved. We refer to a recent survey [TTM*22] for progress on editing neural representations.

Real-Time Performance. Some category-specific methods [TZG*18] are already capable of real-time performance. In the general setting, real-time performance comes at the cost of noticeably lower quality [YRCA15]. Related single-camera settings pave promising paths towards real-time high-quality general dynamic reconstruction: general dynamic RGB-D reconstruction [NFS15, LZYX22] has a long history of real-time speed, and classical sparse RGB SLAM [CEG*21] and neural dense RGB-D SLAM [ZPL*22] also run at real-time rates, with neural dense RGB SLAM very recently achieving the same [CTH*22, RLC22].

Data Bias. Reducing biases in the data is an open challenge not only in 3D reconstruction but in computer vision in general. Different ethnic groups and minorities are underrepresented in most existing datasets, which makes them unbalanced. As a result, methods trained on them (*e.g.* to estimate texture or albedo) are often biased towards statistically expected skin colors (*i.e.* light tones). Special care should be taken when acquiring data so that as many ethnicities as possible are represented in the samples [QWM*20]. Moreover, benchmarks that quantify biases are of great help [FBT*22].

Model Variety. Morphable and parametric models assume able-bodied individuals; missing limbs are seldom modeled. The same holds for highly individualistic appearance variations like tattoos.

Event Cameras. As monocular non-rigid 3D reconstruction from event cameras is an emerging domain with only a few published works, we discuss them jointly here instead of in their respective sections. Their design is challenging as existing RGB-based techniques are not directly applicable to event streams. Event cameras provide an ultra-high temporal event resolution ($\approx 1\mu s$) and record with high dynamic range (see Sec. 2.4). Hence, they are well-suited for high-speed motions in challenging lighting conditions.

EventCap [XXG*20] tracks a human in 3D from a hybrid input of events and synchronous greyscale images captured at $\leq$25 fps. For its highest accuracy, EventCap requires a rigged and skinned 3D human template but also supports SMPL [LMR*15]. Note that the events and images are captured by the same sensor, *i.e. the scene is observed from a single view.* EventCap uses events to track 2D features and establish correspondences between the greyscale keyframes. That is because, for high-speed motions, the 2D point trajectories guided by events can differ significantly from linear feature interpolation between the keyframes. EventHPE [ZGZ*21] relies only on a single greyscale frame for the 3D human pose initialization. It is a learning-based human-specific approach trained on a new dataset with event streams and corresponding SMPL annotations. It uses an unsupervised warping loss with events-based optical flow. Nehvi *et al.* [NGM*21] track general objects from an event camera. Their analysis-by-synthesis SfT approach searches for 3D states obeying the deformation model (such as ARAP or a parametric shape model [RTB17]) and inducing synthetic events that resemble the observed events. The data term accumulates the events into *event frames*, a 2D representation of accumulated events in short time intervals. EventHands [RGW*21] is a data-driven approach for 3D hand pose estimation from a single event stream trained with a synthetic dataset. It neither uses greyscale images nor a 3D template. Both EventHPE and EventHands use parametric models [LMR*15, RTB17]. EventHands enables the tracking of high-speed hand movements at 1000 equivalent fps, *i.e.* the number of discretely reconstructed 3D shapes per second. A Kalman filter stabilizes the results via temporal smoothness.

Observations. EventHands demonstrates that events are more abstract signals than RGB or greyscale pixel values. Thus, the model trained on synthetic data generalizes well to real events. Furthermore, all discussed methods show that high-speed motions could be reconstructed using much lower bandwidth compared to high-speed RGB recordings. Furthermore, they all convert the raw event streams to more suitable 2D representations. Finally, a single or a few events are not expressive enough; a critical mass of events is necessary to regress changes in the estimated 3D poses and shapes.

Physics. Physically-based simulation of soft body dynamics [BMM17] has been well and actively studied for more than 30 years [TPBF87]. However, unlike this well-posed forward problem, non-rigid reconstruction is inverse and ill-posed, and physics-based methods only form an emerging field. In addition to higher computational load, physics-based models are also harder to optimize in practice. For example, physically meaningful material parameters (which determine the deformations) are often time-invariant, which leads to hard-to-escape local minima due to the strong path dependence in the forward simulation. Therefore, physics-based reconstruction methods primarily target simple objects and only elastic phenomena. They ignore complex physics such as human skin, muscles, hair and clothing, which are all non-rigid but have different physical properties and varied deformation behavior. They also do not account for collisions, contacts, fractures, or plasticity.

Along with that, non-learning methods that model physics and data-driven learning-based methods have shown first success. The former necessarily employ some intuition or approximation of physics, and they primarily differ by how accurately they model the physics of deformable objects. Thus, a few earlier SfT and NRSfM methods apply continuum mechanics as hard constraints by representing surfaces and tracking deformations with finite elements (FEM) [MH17] or particle-based models [AMN15, ÖB17]. Recent advances in differentiable simulation [LDW*22, LLK19] and differentiable rendering (Sec. 2.4) enable physics-based analysis by synthesis: ϕ-SfT [KTE*22] reconstructs 3D geometry while others [JBH20, MMG*21] consider the inverse elasticity problem.

While early methods were mostly physics-inspired, the community shifted towards learning, particularly neural networks, in the last decade. Recently, there has been growing interest in combining physics and learning approaches to achieve robust solutions, as physics is the intuition representing invariant properties of the physical world; and learning could extend this world rather than starting from scratch. In the sparse setting, Shimada *et al.* [SGX*21] achieve state-of-the-art results in human motion capture with their *physionical* method, a neural approach that is aware of physical and environmental constraints. However, extending it to the dense case is not trivial; one of the reasons is the increased requirement for computational resources. Neural approaches [RPK19, CRBD18] could additionally aid in solving deformation PDEs that, otherwise, require numerical methods and are often computationally prohibitive. Apart from these methods, there remains a wide range of problems where physics-based solutions remain underexplored, as they have only recently become feasible.

5. Social Implications

We discuss a wide range of potential upsides of monocular reconstruction in the introduction (Sec. 1), like VR/AR, content creation, robotics, medicine, and many others. There are, however, also some potential social downsides, which we discuss in detail here.

Environment. The rise of neural methods increased the usage of GPUs, which can be harmful to the environment and climate due to the material needed, production, and energy usage when running. While the last issue can be addressed by the end user via clean energy, material sourcing and production are more challenging. Still, reconstruction methods that are easy to use might help in environmental research, *e.g.* 3D glacier reconstruction [PRS*16, STC21], thereby ultimately positively impacting the environment.

Privacy and Consent. Easy-to-employ reconstruction methods can potentially lead to unwanted misuse of personal likenesses. Especially when handling datasets containing identifiable data of humans, privacy and consent should be considered, both for training data and at test time when reconstructing other people. Furthermore, editability [KGT*18, YSL*22, MLS*22], which is not a focus of this STAR, could lead to issues with visual content modified or generated with malevolent intent (*e.g.* misinformation). The detection of edited content is an active research area [RCV*19, SY22]. Such detectors often exploit expert knowledge about the design of state-of-the-art methods that generate such content in the first place, which makes continued research necessary. For a discussion specific to neural rendering, we refer to a recent survey [TTM*22].

Inclusiveness. Reconstruction methods can serve as a more inclusive basis for AR/VR if they cover a wider range of variation among people. We refer to *Data Bias* and *Model Variety* in Sec. 4.

Authoritativeness. In certain restricted settings like faces, reconstruction methods are reliable and can help with the virtual ageing of crime victims or face reconstruction from dry skulls [EST*20]. General reconstruction methods, however, should not be taken as authoritative, *e.g.* in legal contexts. Since their problem setting is severely ill-posed, the results are only plausible: consistent but merely possible. They do not infer reliable information about reality beyond what is in the input (*e.g.* how a suspect handled a gun hidden behind their back while being recorded from the front).

Accessibility. The research field is quite accessible: Papers are mirrored on public sites; code and dataset releases are common; some limited GPU resources are accessible for free in the cloud, with larger resources requiring ‘only’ money and no longer one’s own physical infrastructure; and RGB cameras are easily obtainable.

6. Conclusions

We traced how the deep learning revolution, including differentiable rendering and neural rendering, has spread the general non-rigid 3D reconstruction field beyond NRSfM and SfT. This is especially promising considering the saturation of improvements in NRSfM. Still, general methods remain in an early phase and far from being solved, with much of the design space under-explored at best. Category-specific methods for humans and faces are maturing, with close to photorealistic results, while hands and animals, with their unique challenges, have seen comparatively less work. Orthogonal to these, we discussed several promising developments that have recently become practically feasible for reconstruction, like physics simulation and event cameras. In addition, we described the components of the reconstruction pipeline in detail and commented on several open challenges and social implications, which we believe future research would benefit from considering. We hope this STAR will serve as an informative overview for established researchers and a helpful starting point for newcomers entering this exciting and fast-changing area.

7. Acknowledgements.

N. Kairanda, V. Golyanik, and C. Theobalt are supported in part by the ERC Consolidator Grant *4DReply* (770784). A. Kortylewski acknowledges support via his Emmy Noether Research Group funded by the German Science Foundation (DFG) under Grant No. 468670075. The authors thank Rhaleb Zayer for his helpful feedback on Sec. 2. Open Access funding enabled and organized by Projekt DEAL.

References

[AFS*11] Agarwal S., Furukawa Y., Snavely N., Simon I., Curless B., Seitz S. M., Szeliski R.: Building rome in a day. *Communications of the ACM* (2011). 1, 2

[AGS17] Ansari M. D., Golyanik V., Stricker D.: Scalable dense monocular surface reconstruction. In *International Conference on 3D Vision (3DV)* (2017). 12, 13, 14

[AMB*19] ALLDIECK T., MAGNOR M., BHATNAGAR B. L., THEOBALT C., PONS-MOLL G.: Learning to reconstruct people in clothing from a single RGB camera. In *Computer Vision and Pattern Recognition (CVPR)* (2019). 18

[AMN15] AGUDO A., MORENO-NOGUER F.: Simultaneous pose and non-rigid shape with particle dynamics. In *Computer Vision and Pattern Recognition (CVPR)* (2015). 26

[AMN18] AGUDO A., MORENO-NOGUER F.: A scalable, efficient, and accurate solution to non-rigid structure from motion. *Computer Vision and Image Understanding* (2018). 14

[AMNCM16] AGUDO A., MORENO-NOGUER F., CALVO B., MONTIEL J. M. M.: Sequential non-rigid structure from motion using physical priors. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2016). 14

[APMTM19] ALLDIECK T., PONS-MOLL G., THEOBALT C., MAGNOR M.: Tex2shape: Detailed full human body geometry from a single image. In *International Conference on Computer Vision (ICCV)* (2019). 18

[ASKK08] AKHTER I., SHEIKH Y., KHAN S., KANADE T.: Nonrigid structure from motion in trajectory space. In *Advances in Neural Information Processing Systems (NeurIPS)* (2008). 13

[AXS21] ALLDIECK T., XU H., SMINCHISESCU C.: imghum: Implicit generative models of 3d human shape and articulated pose. In *International Conference on Computer Vision (ICCV)* (2021). 18

[AXS*22] ATHAR S., XU Z., SUNKAVALLI K., SHECHTMAN E., SHU Z.: Rignerf: Fully controllable neural 3d portraits. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 20, 22

[AZS22] ALLDIECK T., ZANFIR M., SMINCHISESCU C.: Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 18

[BBC*20] BIGGS B., BOYNE O., CHARLES J., FITZGIBBON A., CIPOLLA R.: Who left the dogs out: 3D animal reconstruction with expectation maximization in the loop. In *European Conference on Computer Vision (ECCV)* (2020). 24

[BBGB16] BÉRARD P., BRADLEY D., GROSS M., BEELER T.: Lightweight eye capture using a parametric model. *ACM Transactions on Graphics* (2016). 22

[BBH14] BRUNET F., BARTOLI A., HARTLEY R. I.: Monocular template-based 3d surface reconstruction: Convex inextensible and nonconvex isometric methods. *Computer Vision and Image Understanding* (2014). 11

[BBT19] BOUKHAYMA A., BEM R. D., TORR P. H.: 3d hand shape and pose from images in the wild. In *Computer Vision and Pattern Recognition (CVPR)* (2019). 23

[BGC*15] BARTOLI A., GÉRARD Y., CHADEBECQ F., COLLINS T., PIZARRO D.: Shape-from-template. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2015). 10, 11, 12

[BHB00] BREGLER C., HERTZMANN A., BIERMANN H.: Recovering non-rigid 3d shape from image streams. In *Computer Vision and Pattern Recognition (CVPR)* (2000). 7, 12, 13

[BHB*11] BEELER T., HAHN F., BRADLEY D., BICKEL B., BEARDSLEY P., GOTSMAN C., SUMNER R. W., GROSS M.: High-quality passive facial performance capture using anchor frames. *ACM Transactions on Graphics* (2011). 14

[BKK19] BAEK S., KIM K. I., KIM T.-K.: Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Computer Vision and Pattern Recognition (CVPR)* (2019). 23

[BMM17] BENDER J., MÜLLER M., MACKLIN M.: A survey on position based dynamics. *Proceedings of the European Association for Computer Graphics: Tutorials* (2017). 25

[Bra05] BRAND M.: A direct method for 3d factorization of nonrigid motion observed in 2d. In *Computer Vision and Pattern Recognition (CVPR)* (2005). 12

[BRFC18] BIGGS B., RODDICK T., FITZGIBBON A., CIPOLLA R.: Creatures great and SMAL: Recovering the shape and motion of animals from video. In *Asian Conference on Computer Vision (ACCV)* (2018). 24

[Bro66] BROWN D. C.: Decentering distortion of lenses. In *Photogrammetric Engineering* (1966). 7

[BTP*22] B R M., TEWARI A., PAN X., ELGHARIB M., THEOBALT C.: gCoRF: Generative compositional radiance fields. In *International Conference on 3D Vision (3DV)* (2022). 21, 22

[BTS*21] B R M., TEWARI A., SEIDEL H.-P., ELGHARIB M., THEOBALT C.: Learning complete 3d morphable face models from images and videos. In *Computer Vision and Pattern Recognition (CVPR)* (2021). 22

[BTTPM19] BHATNAGAR B. L., TIWARI G., THEOBALT C., PONS-MOLL G.: Multi-garment net: Learning to dress 3d people from images. In *International Conference on Computer Vision (ICCV)* (2019). 18

[BV99] BLANZ V., VETTER T.: A morphable model for the synthesis of 3d faces. In *ACM Transactions on Graphics* (1999). 10, 20

[BWM*20] BADGER M., WANG Y., MODH A., PERKES A., KOLOTOUROS N., PFROMMER B., SCHMIDT M., DANIILIDIS K.: 3D bird reconstruction: a dataset, model, and shape recovery from a single view. In *European Conference on Computer Vision (ECCV)* (2020). 24

[CBBC16] COLLINS T., BARTOLI A., BOURDEL N., CANIS M.: Robust, real-time, dense and deformable 3d organ tracking in laparoscopic videos. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2016). 12

[CEG*21] CAMPOS C., ELVIRA R., GOMEZ J. J., MONTIEL J. M. M., TARDOS J. D.: ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. *IEEE Transactions on Robotics* (2021). 25

[CET98] COOTES T. F., EDWARDS G. J., TAYLOR C. J.: Active appearance models. In *European Conference on Computer Vision (ECCV)* (1998). 20

[CF13] CASHMAN T. J., FITZGIBBON A. W.: What shape are dolphins? building 3d morphable models from 2d images. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2013). 24

[CFG*15] CHANG A. X., FUNKHOUSER T., GUIBAS L., HANRAHAN P., HUANG Q., LI Z., SAVARESE S., SAVVA M., SONG S., SU H., ET AL.: Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015). 4

[CHV*22] CORONA E., HODAN T., VO M., MORENO-NOGUER F., SWEENEY C., NEWCOMBE R., MA L.: Lisa: Learning implicit shape and appearance of hands. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 23

[CLC*22] CHAN E. R., LIN C. Z., CHAN M. A., NAGANO K., PAN B., MELLO S. D., GALLO O., GUIBAS L., TREMBLAY J., KHAMIS S., KARRAS T., WETZSTEIN G.: Efficient geometry-aware 3D generative adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 1, 19, 20, 21, 22

[CMK*21] CHAN E., MONTEIRO M., KELLNHOFER P., WU J., WETZSTEIN G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Computer Vision and Pattern Recognition (CVPR)* (2021). 20, 21, 22

[COR*16] CORDTS M., OMRAN M., RAMOS S., REHFELD T., ENZWEILER M., BENENSON R., FRANKE U., ROTH S., SCHIELE B.: The cityscapes dataset for semantic urban scene understanding. In *Computer Vision and Pattern Recognition (CVPR)* (2016). 2

[CPA*21] CORONA E., PUMAROLA A., ALENYÀ G., PONS-MOLL G., MORENO-NOGUER F.: Smplicit: Topology-aware generative model for clothed people. In *Computer Vision and Pattern Recognition (CVPR)* (2021). 18

[CPBC16] CHHATKULI A., PIZARRO D., BARTOLI A., COLLINS T.: A stable analytical framework for isometric shape-from-template by surface integration. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2016). 11, 12

[CPPFJ*19] Casillas-Perez D., Pizarro D., Fuentes-Jimenez D., Mazo M., Bartoli A.: Equiareal shape-from-template. *Journal of Mathematical Imaging and Vision* (2019). 11, 12

[CPPFJ*21] Casillas-Perez D., Pizarro D., Fuentes-Jimenez D., Mazo M., Bartoli A.: The isowarp: the template-based visual geometry of isometric surfaces. *International Journal of Computer Vision (IJCV)* (2021). 11, 12

[CRBD18] Chen R. T. Q., Rubanova Y., Bettencourt J., Duvenaud D.: Neural ordinary differential equations. *Advances in Neural Information Processing Systems (NeurIPS)* (2018). 15, 26

[CRZ*22] Cai Z., Ren D., Zeng A., Lin Z., Yu T., Wang W., Fan X., Gao Y., Yu Y., Pan L., Hong F., Zhang M., Loy C. C., Yang L., Liu Z.: Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision (ECCV)* (2022). 19

[CSW*16] Chai M., Shao T., Wu H., Weng Y., Zhou K.: Autohair: fully automatic hair modeling from a single image. *ACM Transactions on Graphics* (2016). 22

[CTH*22] Chung C.-M., Tseng Y.-C., Hsu Y.-C., Shi X.-Q., Hua Y.-H., Yeh J.-F., Chen W.-C., Chen Y.-T., Hsu W. H.: Orbeez-slam: A real-time monocular visual slam with orb features and nerf-realized mapping. *arXiv preprint* (2022). 25

[CTM*21] Caron M., Touvron H., Misra I., Jégou H., Mairal J., Bojanowski P., Joulin A.: Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)* (2021). 8

[CYX*21] Chao Y.-W., Yang W., Xiang Y., Molchanov P., Handa A., Tremblay J., Narang Y. S., Van Wyk K., Iqbal U., Birchfield S., Kautz J., Fox D.: Dexycb: A benchmark for capturing hand grasping of objects. In *Computer Vision and Pattern Recognition (CVPR)* (2021). 24

[CZ19] Chen Z., Zhang H.: Learning implicit fields for generative shape modeling. *Computer Vision and Pattern Recognition (CVPR)* (2019). 6

[CZB*21] Chen X., Zheng Y., Black M. J., Hilliges O., Geiger A.: Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *International Conference on Computer Vision (ICCV)* (2021). 7

[DC16] Do Carmo M. P.: *Differential geometry of curves and surfaces: revised and updated second edition*. Courier Dover Publications, 2016. 4

[DLH12] Dai Y., Li H., He M.: A simple prior-free method for non-rigid structure-from-motion factorization. In *Computer Vision and Pattern Recognition (CVPR)* (2012). 12, 13

[DP22] Duggal S., Pathak D.: Topologically-aware deformation fields for single-view 3d reconstruction. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 17

[DPS18] Dai H., Pears N., Smith W.: A data-augmented 3d morphable model of the ear. In *Proc. International Conference on Automatic Face and Gesture Recognition* (2018). 22

[DYXT22] Deng Y., Yang J., Xiang J., Tong X.: Gram: Generative radiance manifolds for 3d-aware image generation. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 20, 21, 22

[DZY*21] Du Y., Zhang Y., Yu H.-X., Tenenbaum J. B., Wu J.: Neural radiance flow for 4d view synthesis and video processing. In *International Conference on Computer Vision (ICCV)* (2021). 15

[EÖC*20] Espinel Y., Özgür E., Calvet L., Le Roy B., Buc E., Bartoli A.: Combining visual cues with interactions for 3d–2d registration in liver laparoscopy. *Annals of Biomedical Engineering* (2020). 12

[EST*20] Egger B., Smith W. A., Tewari A., Wuhrer S., Zollhoefer M., Beeler T., Bernard F., Bolkart T., Kortylewski A., Romdhani S., et al.: 3d morphable face models—past, present, and future. *ACM Transactions on Graphics* (2020). 3, 20, 21, 22, 26

[FBT*22] Feng H., Bolkart T., Tesch J., Black M. J., Abrevaya V.: Towards racially unbiased skin tone estimation via scene disambiguation. In *European Conference on Computer Vision (ECCV)* (2022). 1, 20, 25

[FFBB21] Feng Y., Feng H., Black M. J., Bolkart T.: Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* (2021). 20, 22

[FJCPP*18] Fuentes-Jimenez D., Casillas-Perez D., Pizarro D., Collins T., Bartoli A.: Deep shape-from-template: Wide-baseline, dense and fast registration and deformable reconstruction from a single image. *arXiv preprint arXiv:1811.07791* (2018). 11

[FJPCP*21] Fuentes-Jimenez D., Pizarro D., Casillas-Perez D., Collins T., Bartoli A.: Texture-generic deep shape-from-template. *IEEE Access* (2021). 11, 12

[FRA11] Fayad J., Russell C., Agapito L.: Automated articulated structure and 3d shape recovery from point correspondences. In *International Conference on Computer Vision (ICCV)* (2011). 11

[FVVD*96] Foley J. D., Van F. D., Van Dam A., Feiner S. K., Hughes J. F.: *Computer graphics: principles and practice*. Addison-Wesley Professional, 1996. 3

[FYW*22] Fang J., Yi T., Wang X., Xie L., Zhang X., Liu W., Niessner M., Tian Q.: Fast dynamic radiance fields with time-aware neural voxels. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* (2022). 15

[GB22] Grasshof S., Brandt S. S.: Tensor-based non-rigid structure from motion. In *Winter Conference on Applications of Computer Vision (WACV)* (2022). 13, 14

[GBCR16] Garg R., BG V. K., Carneiro G., Reid I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision (ECCV)* (2016). 2

[GCD*22] Guo X., Chen G., Dai Y., Ye X., Sun J., Tan X., Ding E.: Neural deformable voxel grid for fast optimization of dynamic view synthesis. In *Asian Conference on Computer Vision (ACCV)* (2022). 15

[GDO*22] Gallego G., Delbruck T., Orchard G., Bartolozzi C., Taba B., Censi A., Leutenegger S., Davison A. J., Conradt J., Daniilidis K., Scaramuzza D.: Event-based vision: A survey. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2022). 3, 6

[GFK*18] Groueix T., Fisher M., Kim V. G., Russell B., Aubry M.: AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *Computer Vision and Pattern Recognition (CVPR)* (2018). 6

[GFM*19] Gabeur V., Franco J.-S., Martin X., Schmid C., Rogez G.: Moulding humans: Non-parametric 3d human shape estimation from single images. In *International Conference on Computer Vision (ICCV)* (2019). 17

[GFS17a] Golyanik V., Fetzer T., Stricker D.: Accurate 3d reconstruction of dynamic scenes from monocular image sequences with severe occlusions. In *Winter Conference on Applications of Computer Vision (WACV)* (2017). 13, 14

[GFS17b] Golyanik V., Fetzer T., Stricker D.: Introduction to coherent depth fields for dense monocular surface recovery. In *British Machine Vision Conference (BMVC)* (2017). 14

[GJS19] Golyanik V., Jonas A., Stricker D.: Consolidating segmentwise non-rigid structure from motion. In *Machine Vision Applications (MVA)* (2019). 14

[GJST20] Golyanik V., Jonas A., Stricker D., Theobalt C.: Intrinsic dynamic shape prior for dense non-rigid structure from motion. In *International Conference on 3D Vision (3DV)* (2020). 13, 14

[GKM20] Goel S., Kanazawa A., , Malik J.: Shape and viewpoints without keypoints. In *European Conference on Computer Vision (ECCV)* (2020). 16

[GLT*22] Gao H., Li R., Tulsiani S., Russell B., Kanazawa A.: Monocular dynamic view synthesis: A reality check. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022). 15

[GLWT22] Gu J., Liu L., Wang P., Theobalt C.: Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)* (2022). 20, 21, 22

[GM11] Gotardo P. F. U., Martinez A. M.: Kernel non-rigid structure from motion. In *International Conference on Computer Vision (ICCV)* (2011). 12

[GPCB20] Gallardo M., Pizarro D., Collins T., Bartoli A.: Shape-from-template with curves. *International Journal of Computer Vision (IJCV)* (2020). 11

[GPL*22] Grassal P.-W., Prinzler M., Leistner T., Rother C., Niessner M., Thies J.: Neural head avatars from monocular rgb videos. *Computer Vision and Pattern Recognition (CVPR)* (2022). 20, 22

[GRA13a] Garg R., Roussos A., Agapito L.: Dense variational reconstruction of non-rigid surfaces from monocular video. In *Computer Vision and Pattern Recognition (CVPR)* (2013). 12, 13, 14

[GRA13b] Garg R., Roussos A., Agapito L.: A variational approach to video registration with subspace constraints. *International Journal of Computer Vision (IJCV)* (2013). 12, 14

[GRL*19] Ge L., Ren Z., Li Y., Xue Z., Wang Y., Cai J., Yuan J.: 3d hand shape and pose estimation from a single rgb image. In *Computer Vision and Pattern Recognition (CVPR)* (2019). 23

[GSKH21] Gao C., Saraf A., Kopf J., Huang J.-B.: Dynamic view synthesis from dynamic monocular video. In *International Conference on Computer Vision (ICCV)* (2021). 15

[GSVS18] Golyanik V., Shimada S., Varanasi K., Stricker D.: Hdm-net: Monocular non-rigid 3d reconstruction with learned deformation model. In *EuroVR* (2018). 11, 12

[GTZN21] Gafni G., Thies J., Zollhöfer M., Niessner M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Computer Vision and Pattern Recognition (CVPR)* (2021). 20, 22

[GZC*16] Garrido P., Zollhöfer M., Casas D., Valgaerts L., Varanasi K., Pérez P., Theobalt C.: Reconstruction of personalized 3d face rigs from monocular video. *ACM Transactions on Graphics* (2016). 2

[HBAD21] Hoppe Nesgaard Jensen S., Brix Doest M. E., Aanaes H., Del Bue A.: A Benchmark and Evaluation of Non-Rigid Structure from Motion. *International Journal of Computer Vision (IJCV)* (2021). 2, 4, 14

[HC17] Haouchine N., Cotin S.: Template-based monocular 3d recovery of elastic shapes using lagrangian multipliers. In *Computer Vision and Pattern Recognition (CVPR)* (2017). 11

[HCJS20] He T., Collomosse J., Jin H., Soatto S.: Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)* (2020). 17

[HPX*22] Hong Y., Peng B., Xiao H., Liu L., Zhang J.: Headnerf: A real-time nerf-based parametric head model. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 19, 20, 21, 22

[HSRL22] Hampali S., Sarkar S. D., Rad M., Lepetit V.: Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 24

[HSW*17] Hu L., Saito S., Wei L., Nagano K., Seo J., Fursund J., Sadeghi I., Sun C., Chen Y.-C., Li H.: Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics* (2017). 22

[HTB*20] Hasson Y., Tekin B., Bogo F., Laptev I., Pollefeys M., Schmid C.: Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Computer Vision and Pattern Recognition (CVPR)* (2020). 24

[HVT*19] Hasson Y., Varol G., Tzionas D., Kalevatykh I., Black M. J., Laptev I., Schmid C.: Learning joint reconstruction of hands and manipulated objects. In *Computer Vision and Pattern Recognition (CVPR)* (2019). 24

[HXL*20] Huang Z., Xu Y., Lassner C., Li H., Tung T.: Arch: Animatable reconstruction of clothed humans. In *Computer Vision and Pattern Recognition (CVPR)* (2020). 18

[HXR*18] Habermann M., Xu W., Rhodin H., Zollhöfer M., Pons-Moll G., Theobalt C.: Nrst: Non-rigid surface tracking from monocular video. In *German Conference for Pattern Recognition (GCPR)* (2018). 11, 12

[HXS*21] He T., Xu Y., Saito S., Soatto S., Tung T.: Arch++: Animation-ready clothed human reconstruction revisited. In *International Conference on Computer Vision (ICCV)* (2021). 18

[HXZ*19] Habermann M., Xu W., Zollhöfer M., Pons-Moll G., Theobalt C.: Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics* (2019). 18, 19

[HXZ*20] Habermann M., Xu W., Zollhoefer M., Pons-Moll G., Theobalt C.: Deepcap: Monocular human performance capture using weak supervision. In *Computer Vision and Pattern Recognition (CVPR)* (2020). 1, 4, 19

[HYH*20] Heming Z., Yu C., Hang J., Weikai C., Dong D., Zhangye W., Shuguang C., Xiaoguang H.: Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *European Conference on Computer Vision (ECCV)* (2020). 19

[ILBH*11] Indiveri G., Linares-Barranco B., Hamilton T., van Schaik A., Etienne-Cummings R., Delbruck T., Liu S.-C., Dudek P., Häfliger P., Renaud S., Schemmel J., Cauwenberghs G., Arthur J., Hynna K., Folowosele F., Saïghi S., Serrano-Gotarredona T., Wijekoon J., Wang Y., Boahen K.: Neuromorphic silicon neuron circuits. *Frontiers in Neuroscience* (2011). 6

[JBH20] Jaques M., Burke M., Hospedales T.: Physics-as-inverse-graphics: Unsupervised physical parameter estimation from video. In *International Conference on Learning Representations (ICLR)* (2020). 26

[JCSN20] Jinka S., Chacko R., Sharma A., Narayanan P.: Peeledhuman: Robust shape representation for textured 3d human body reconstruction. In *International Conference on 3D Vision (3DV)* (2020). 17

[JHGT22] Jiang Y., Habermann M., Golyanik V., Theobalt C.: Hifecap: Monocular high-fidelity and expressive capture of human performances. In *British Machine Vision Conference (BMVC)* (2022). 19

[JHS*22] Johnson E. C., Habermann M., Shimada S., Golyanik V., Theobalt C.: Unbiased 4d: Monocular 4d reconstruction with a neural deformation model. *arXiv:2206.08368* (2022). 1, 14, 15

[JYS*22] Jiang W., Yi K. M., Samei G., Tuzel O., Ranjan A.: Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision (ECCV)* (2022). 18, 25

[KB15] Kingma D. P., Ba J.: Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)* (2015). 11

[KCDL18] Kumar S., Cherian A., Dai Y., Li H.: Scalable dense non-rigid structure-from-motion: A grassmannian perspective. In *Computer Vision and Pattern Recognition (CVPR)* (2018). 12, 13, 14

[KDL16] Kumar S., Dai Y., Li H.: Multi-body non-rigid structure-from-motion. In *International Conference on 3D Vision (3DV)* (2016). 13

[KGFT20] Kulkarni N., Gupta A., Fouhey D. F., Tulsiani S.: Articulation-aware canonical surface mapping. In *Computer Vision and Pattern Recognition (CVPR)* (2020). 16

[KGT*18] Kim H., Garrido P., Tewari A., Xu W., Thies J., Niessner M., Pérez P., Richardt C., Zollöfer M., Theobalt C.: Deep video portraits. *ACM Transactions on Graphics* (2018). 26

[KGT19] Kulkarni N., Gupta A., Tulsiani S.: Canonical surface mapping via geometric cycle consistency. In *International Conference on Computer Vision (ICCV)* (2019). 16

[KK21a] Kokkinos F., Kokkinos I.: Learning monocular 3d reconstruction of articulated categories from motion. In *Computer Vision and Pattern Recognition (CVPR)* (2021). 17

[KK21b] Kokkinos F., Kokkinos I.: To the point: Correspondence-driven monocular 3d category reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)* (2021). 17

[KL16] Kong C., Lucey S.: Prior-less compressible structure from motion. In *Computer Vision and Pattern Recognition (CVPR)* (2016). 12

[KLA19] Karras T., Laine S., Aila T.: A style-based generator architecture for generative adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)* (2019). 2, 21

[KLA*20] Karras T., Laine S., Aittala M., Hellsten J., Lehtinen J., Aila T.: Analyzing and improving the image quality of StyleGAN. In *Computer Vision and Pattern Recognition (CVPR)* (2020). 21

[KÖR*17] Koo B., Özgür E., Roy B. L., Buc E., Bartoli A.: Deformable registration of a preoperative 3d liver volume to a laparoscopy image using contour and shading cues. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2017). 12

[KPL*22] Khan M. S. U., Pagani A., Liwicki M., Stricker D., Afzal M. Z.: 3d reconstruction from a single rgb image using deep learning: A review. *Journal of Imaging* (2022). 2

[KTE*22] Kairanda N., Tretschk E., Elgharib M., Theobalt C., Golyanik V.: ϕ-sft: Shape-from-template with a physics-based deformation model. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 1, 2, 6, 11, 12, 26

[KTEM18] Kanazawa A., Tulsiani S., Efros A. A., Malik J.: Learning category-specific mesh reconstruction from image collections. In *European Conference on Computer Vision (ECCV)* (2018). 2, 7, 16, 24

[KTS*21] Kwon T., Tekin B., Stühmer J., Bogo F., Pollefeys M.: H2o: Two hands manipulating objects for first person interaction recognition. In *International Conference on Computer Vision (ICCV)* (2021). 24

[KUH18] Kato H., Ushiku Y., Harada T.: Neural 3d mesh renderer. In *Computer Vision and Pattern Recognition (CVPR)* (2018). 8

[Kum19] Kumar S.: Jumping manifolds: Geometry aware dense non-rigid structure from motion. In *Computer Vision and Pattern Recognition (CVPR)* (2019). 13, 14

[Kum20] Kumar S.: Non-rigid structure from motion: Prior-free factorization method revisited. In *Winter Conference on Applications of Computer Vision (WACV)* (2020). 13

[KYZ*20] Karunratanakul K., Yang J., Zhang Y., Black M., Muandet K., Tang S.: Grasping field: Learning implicit representations for human grasps. In *International Conference on 3D Vision (3DV)* (2020). 24

[Lar22] Laraudogoitia J. P.: Undeformable bodies that are not rigid bodies: A philosophical journey through some (unexpected) supertasks. *Axiomathes* (2022). 2

[LAZ*22] Li M., An L., Zhang H., Wu L., Chen F., Yu T., Liu Y.: Interacting attention graph for single image two-hand reconstruction. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 24

[LB14] Loper M. M., Black M. J.: Opendr: An approximate differentiable renderer. In *European Conference on Computer Vision (ECCV)* (2014). 8

[LC87] Lorensen W. E., Cline H. E.: Marching cubes: A high resolution 3d surface construction algorithm. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* (1987). 6

[LDG18] Liao Y., Donne S., Geiger A.: Deep marching cubes: Learning explicit surface representations. In *Computer Vision and Pattern Recognition (CVPR)* (2018). 6

[LDW*22] Li Y., Du T., Wu K., Xu J., Matusik W.: Diffcloth: Differentiable cloth simulation with dry frictional contact. *ACM Transactions on Graphics* (2022). 26

[LGC*05] Lensch H. P., Goesele M., Chuang Y.-Y., Hawkins T., Marschner S., Matusik W., Mueller G.: Realistic materials in computer graphics. In *ACM SIGGRAPH Courses*. 2005. 5

[LH87] Longuet-Higgins H. C.: A computer algorithm for reconstructing a scene from two projections. *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms* (1987). 1

[LHM*18] Liang S., Huang X., Meng X., Chen K., Shapiro L. G., Kemelmacher-Shlizerman I.: Video to Fully Automatic 3D Hair Model. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* (2018). 22

[LHS*20] Luo X., Huang J., Szeliski R., Matzen K., Kopf J.: Consistent video depth estimation. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* (2020). 15

[LHT*21] Li Y., Habermann M., Thomaszewski B., Coros S., Beeler T., Theobalt C.: Deep Physics-aware Inference of Cloth Deformation for Monocular Human Performance Capture. In *International Conference on 3D Vision (3DV)* (2021). 19

[LL21] Li C., Lee G. H.: Coarse-to-fine animal pose and shape estimation. In *Advances in Neural Information Processing Systems (NeurIPS)* (2021). 2, 24

[LLCL19] Liu S., Li T., Chen W., Li H.: Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *International Conference on Computer Vision (ICCV)* (2019). 8

[LLDM*20] Li X., Liu S., De Mello S., Kim K., Wang X., Yang M.-H., Kautz J.: Online adaptation for consistent mesh reconstruction in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)* (2020). 16

[LLK19] Liang J., Lin M., Koltun V.: Differentiable cloth simulation for inverse problems. In *Advances in Neural Information Processing Systems (NeurIPS)* (2019). 12, 26

[LLK*20] Li X., Liu S., Kim K., De Mello S., Jampani V., Yang M.-H., Kautz J.: Self-supervised single-view 3d reconstruction via semantic consistency. In *European Conference on Computer Vision (ECCV)* (2020). 16

[LLWT15] Liu Z., Luo P., Wang X., Tang X.: Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)* (2015). 2

[LMFV*23] Li C., Morel-Forster A., Vetter T., Egger B., Kortylewski A.: To fit or not to fit: Model-based face reconstruction and occlusion segmentation from weak supervision. In *Computer Vision and Pattern Recognition (CVPR)* (2023). 19, 20, 22

[LMR*15] Loper M., Mahmood N., Romero J., Pons-Moll G., Black M. J.: SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* (2015). 7, 18, 25

[LNSW21] Li Z., Niklaus S., Snavely N., Wang O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Computer Vision and Pattern Recognition (CVPR)* (2021). 14, 15, 25

[Low04] Lowe D. G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)* (2004). 8

[LSS*19] Lombardi S., Simon T., Saragih J., Schwartz G., Lehrmann A., Sheikh Y.: Neural volumes: Learning dynamic renderable volumes from images. *ACM Transactions on Graphics* (2019). 7, 8

[LXWY22] Li P., Xu Y., Wei Y., Yang Y.: Self-correction for human parsing. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2022). 2

[LYYA*16] Liu-Yin Q., Yu R., Agapito L., Fitzgibbon A., Russell C.: Better together: Joint reasoning for non-rigid 3d reconstruction with specularities and shading. *British Machine Vision Conference (BMVC)* (2016). 11, 12

[LZYX22] Lin W., Zheng C., Yong J.-H., Xu F.: Occlusionfusion: Occlusion-aware motion estimation for real-time dynamic 3d reconstruction. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 25

[MAMT15] Mur-Artal R., Montiel J. M. M., Tardós J. D.: Orb-slam: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics* (2015). 1

[MBH15] Malti A., Bartoli A., Hartley R.: A linear least-squares solution to elastic shape-from-template. In *Computer Vision and Pattern Recognition (CVPR)* (2015). 11, 12

[MCL22] Moon G., Choi H., Lee K. M.: Neuralannot: Neural annotator for 3d human mesh training sets. In *Computer Vision and Pattern Recognition (CVPR) Workshops* (2022). 24

[MH17] Malti A., Herzet C.: Elastic shape-from-template with spatially sparse deforming forces. In *Computer Vision and Pattern Recognition (CVPR)* (2017). 11, 26

[MHBK13] Malti A., Hartley R., Bartoli A., Kim J.-H.: Monocular template-based 3d reconstruction of extensible surfaces with local linear elasticity. In *Computer Vision and Pattern Recognition (CVPR)* (2013). 12

[MJK*22] Mojtaba Marvasti-Zadeh S., Jahromi M. N. S., Khaghani J., Goodsman D., Ray N., Erbilgin N.: Learning-based monocular 3d reconstruction of birds: A contemporary survey. *arXiv e-prints* (2022). 3, 16, 24

[MLS*22] Menapace W., Lathuilière S., Siarohin A., Theobalt C., Tulyakov S., Golyanik V., Ricci E.: Playable environments: Video manipulation in space and time. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 25, 26

[MMG*21] Murthy J. K., Macklin M., Golemo F., Voleti V., Petrini L., Weiss M., Considine B., Parent-Lévesque J., Xie K., Erleben K., Paull L., Shkurti F., Nowrouzezahrai D., Fidler S.: gradsim: Differentiable simulation for system identification and visuomotor control. In *International Conference on Learning Representations (ICLR)* (2021). 26

[MNPF10] Moreno-Noguer F., Porta J. M., Fua P.: Exploring ambiguities for monocular non-rigid shape estimation. In *European Conference on Computer Vision (ECCV)* (2010). 8

[MON*19] Mescheder L., Oechsle M., Niemeyer M., Nowozin S., Geiger A.: Occupancy networks: Learning 3d reconstruction in function space. In *Computer Vision and Pattern Recognition (CVPR)* (2019). 6

[Mor78] Moré J. J.: The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis*. 1978. 11

[MPJ*19] Michalkiewicz M., Pontes J. K., Jack D., Baktashmotlagh M., Eriksson A.: Deep level sets: Implicit surface representations for 3d shape inference. *arXiv preprint arXiv:1901.06802* (2019). 6

[MSL20] Moon G., Shiratori T., Lee K. M.: Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *European Conference on Computer Vision (ECCV)* (2020). 23

[MSL*23] Menapace W., Siarohin A., Lathuilière S., Achlioptas P., Golyanik V., Ricci E., Tulyakov S.: Plotting behind the scenes: Towards learnable game engines. *arXiv e-prints* (2023). 25

[MST*20] 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 *European Conference on Computer Vision (ECCV)* (2020). 2, 4, 6, 7, 8, 21

[MSY10] Mountney P., Stoyanov D., Yang G.-Z.: Three-dimensional tissue deformation recovery and tracking. *IEEE Signal Processing Magazine* (2010). 14

[MYW*20] Moon G., Yu S.-I., Wen H., Shiratori T., Lee K. M.: Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)* (2020). 2, 24

[NFS15] Newcombe R. A., Fox D., Seitz S. M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Computer Vision and Pattern Recognition (CVPR)* (2015). 25

[NGM*21] Nehvi J., Golyanik V., Mueller F., Seidel H.-P., Elgharib M., Theobalt C.: Differentiable event stream simulator for non-rigid 3d tracking. In *Computer Vision and Pattern Recognition (CVPR) Workshops* (2021). 11, 25

[NMOG20] Niemeyer M., Mescheder L., Oechsle M., Geiger A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Computer Vision and Pattern Recognition (CVPR)* (2020). 7, 8

[NNS*20] Neverova N., Novotny D., Szafraniec M., Khalidov V., Labatut P., Vedaldi A.: Continuous surface embeddings. *Advances in Neural Information Processing Systems (NeurIPS)* (2020). 8, 16

[NÖF15] Ngo D. T., Östlund J., Fua P.: Template-based monocular 3d shape recovery using laplacian meshes. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2015). 11

[NPJ*15] Ngo D. T., Park S., Jorstad A., Crivellaro A., Yoo C. D., Fua P.: Dense image registration and deformable surface reconstruction in presence of occlusions and minimal texture. In *International Conference on Computer Vision (ICCV)* (2015). 11

[NRG*19] Novotny D., Ravi N., Graham B., Neverova N., Vedaldi A.: C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In *International Conference on Computer Vision (ICCV)* (2019). 14

[ÖB17] Özgür E., Bartoli A.: Particle-sft: A provably-convergent, fast shape-from-template algorithm. *International Journal of Computer Vision (IJCV)* (2017). 11, 26

[OELS*22] Or-El R., Luo X., Shan M., Shechtman E., Park J. J., Kemelmacher-Shlizerman I.: StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 20, 21, 22

[OMT*21] Ost J., Mannan F., Thuerey N., Knodt J., Heide F.: Neural scene graphs for dynamic scenes. In *Computer Vision and Pattern Recognition (CVPR)* (2021). 25

[PAP*18] Pumarola A., Agudo A., Porzi L., Sanfeliu A., Lepetit V., Moreno-Noguer F.: Geometry-Aware Network for Non-Rigid Shape Prediction from a Single View. In *Computer Vision and Pattern Recognition (CVPR)* (2018). 11, 12

[PCG*19] 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 *Computer Vision and Pattern Recognition (CVPR)* (2019). 2, 18

[PCPMMN21] Pumarola A., Corona E., Pons-Moll G., Moreno-Noguer F.: D-nerf: Neural radiance fields for dynamic scenes. In *Computer Vision and Pattern Recognition (CVPR)* (2021). 15

[PDBX*12] Paladini M., Del Bue A., Xavier J., Agapito L., Stosić M., Dodig M.: Optimal metric projections for deformable and articulated structure-from-motion. *International Journal of Computer Vision (IJCV)* (2012). 12

[PFS*19] PARK J. J., FLORENCE P., STRAUB J., NEWCOMBE R., LOVEGROVE S.: Deepsdf: Learning continuous signed distance functions for shape representation. In *Computer Vision and Pattern Recognition (CVPR)* (2019). 6, 15, 20

[PLK20] PARK S., LEE M., KWAK N.: Procrustean regression networks: Learning 3d structure of non-rigid objects from 2d annotations. In *European Conference on Computer Vision (ECCV)* (2020). 14

[POM*22] PARK J., OH Y., MOON G., CHOI H., LEE K. M.: Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 23

[PPB19] PARASHAR S., PIZARRO D., BARTOLI A.: Local deformable 3d reconstruction with cartan's connections. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2019). 11

[PPBC15] PARASHAR S., PIZARRO D., BARTOLI A., COLLINS T.: As-rigid-as-possible volumetric shape-from-template. In *International Conference on Computer Vision (ICCV)* (2015). 11, 12

[PRS*16] PELLITERO R., REA B. R., SPAGNOLO M., BAKKE J., IVY-OCHS S., FREW C. R., HUGHES P., RIBOLINI A., LUKAS S., RENSSEN H.: Glare, a gis tool to reconstruct the 3d surface of palaeoglaciers. *Computers & Geosciences* (2016). 26

[PSB*21] PARK K., SINHA U., BARRON J. T., BOUAZIZ S., GOLDMAN D. B., SEITZ S. M., MARTIN-BRUALLA R.: Nerfies: Deformable neural radiance fields. In *International Conference on Computer Vision (ICCV)* (2021). 7, 9, 15, 25

[PSF20] PARASHAR S., SALZMANN M., FUA P.: Local non-rigid structure-from-motion from diffeomorphic mappings. In *Computer Vision and Pattern Recognition (CVPR)* (2020). 13, 14

[PSH*21] PARK K., SINHA U., HEDMAN P., BARRON J. T., BOUAZIZ S., GOLDMAN D. B., MARTIN-BRUALLA R., SEITZ S. M.: Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. *ACM Transactions on Graphics* (2021). 15, 17

[PTLT22] PAN X., TEWARI A., LIU L., THEOBALT C.: Gan2x: Non-lambertian inverse rendering of image gans. In *International Conference on 3D Vision (3DV)* (2022). 21, 22

[PVO*20] PLOUMPIS S., VERVERAS E., O'SULLIVAN E., MOSCHOGLOU S., WANG H., PEARS N., SMITH W., GECER B., ZAFEIRIOU S. P.: Towards a complete 3d morphable model of the human head. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2020). 22

[PZX*21] 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 *Computer Vision and Pattern Recognition (CVPR)* (2021). 18

[QGL22] QIAO Y.-L., GAO A., LIN M. C.: Neuphysics: Editable neural geometry and physics from monocular videos. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022). 15

[QWM*20] QIAN N., WANG J., MUELLER F., BERNARD F., GOLYANIK V., THEOBALT C.: HTML: A Parametric Hand Texture Model for 3D Hand Reconstruction and Personalization. In *European Conference on Computer Vision (ECCV)* (2020). 2, 23, 25

[RAG18] RIZA ALP GÜLER NATALIA NEVEROVA I. K.: Densepose: Dense human pose estimation in the wild. In *Computer Vision and Pattern Recognition (CVPR)* (2018). 16

[RCV*19] RÖSSLER A., COZZOLINO D., VERDOLIVA L., RIESS C., THIES J., NIESSNER M.: Faceforensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)* (2019). 26

[RES*22] RUDNEV V., ELGHARIB M., SMITH W., LIU L., GOLYANIK V., THEOBALT C.: Nerf for outdoor scene relighting. In *European Conference on Computer Vision (ECCV)* (2022). 4

[RFA11] RUSSELL C., FAYAD J., AGAPITO L.: Energy based multiple model fitting for non-rigid structure from motion. In *Computer Vision and Pattern Recognition (CVPR)* (2011). 14

[RFA12] RUSSELL C., FAYAD J., AGAPITO L.: Dense non-rigid structure from motion. In *International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission* (2012). 13

[RGW*21] RUDNEV V., GOLYANIK V., WANG J., SEIDEL H.-P., MUELLER F., ELGHARIB M., THEOBALT C.: Eventhands: Real-time neural 3d hand pose estimation from an event stream. In *International Conference on Computer Vision (ICCV)* (2021). 25

[RKH*21] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., ET AL.: Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)* (2021). 24

[RLC22] ROSINOL A., LEONARD J. J., CARLONE L.: Nerf-slam: Real-time dense monocular slam with neural radiance fields. *arXiv preprint* (2022). 25

[RLR*20] REMELLI E., LUKOIANOV A., RICHTER S., GUILLARD B., BAGAUTDINOV T., BAQUE P., FUA P.: Meshsdf: Differentiable iso-surface extraction. In *Advances in Neural Information Processing Systems (NeurIPS)* (2020). 6

[RMY*22] REBAIN D., MATTHEWS M., YI K. M., LAGUN D., TAGLIASACCHI A.: Lolnerf: Learn from one look. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 21, 22

[RPK19] RAISSI M., PERDIKARIS P., KARNIADAKIS G. E.: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics* (2019). 26

[RSJ21] RONG Y., SHIRATORI T., JOO H.: Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *International Conference on Computer Vision (ICCV) Workshops* (2021). 24

[RTB17] ROMERO J., TZIONAS D., BLACK M. J.: Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* (2017). 7, 23, 25

[RTE*21] RAMON E., TRIGINER G., ESCUR J., PUMAROLA A., GARCIA J., GIRO-I NIETO X., MORENO-NOGUER F.: H3d-net: Few-shot high-fidelity 3d head reconstruction. In *International Conference on Computer Vision (ICCV)* (2021). 20, 22

[RZSB22] RÜEGG N., ZUFFI S., SCHINDLER K., BLACK M. J.: BARC: Learning to regress 3d dog shape from images by exploiting breed information. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 24

[SA07] SORKINE O., ALEXA M.: As-rigid-as-possible surface modeling. In *Proc. of Eurographics/ACM SIGGRAPH symposium on Geometry processing* (2007). 10

[SB12] SIFAKIS E., BARBIC J.: Fem simulation of 3d deformable solids: a practitioner's guide to theory, discretization and model reduction. In *ACM SIGGRAPH Courses*. 2012. 4

[SB21] SENGUPTA A., BARTOLI A.: Colonoscopic 3d reconstruction by tubular non-rigid structure-from-motion. *International Journal of Computer Assisted Radiology and Surgery (IJCARS)* (2021). 13

[SBFB19] SANYAL S., BOLKART T., FENG H., BLACK M.: Learning to regress 3D face shape and expression from an image without 3D supervision. In *Computer Vision and Pattern Recognition (CVPR)* (2019). 20

[Sch86] SCHILLER P. H.: The central visual system. *Vision Research* (1986). 1

[SCOL*04] SORKINE O., COHEN-OR D., LIPMAN Y., ALEXA M., RÖSSL C., SEIDEL H.-P.: Laplacian surface editing. In *Proc. of Eurographics/ACM SIGGRAPH symposium on Geometry processing* (2004). 17

[SF10] SALZMANN M., FUA P.: Deformable surface 3d reconstruction from monocular images. *Synthesis Lectures on Computer Vision* (2010). 2, 10

[SGTS19] Shimada S., Golyanik V., Theobalt C., Stricker D.: Ismo-gan: Adversarial learning for monocular non-rigid 3d reconstruction. In *Computer Vision and Pattern Recognition (CVPR) Workshops* (2019). 11, 12

[SGX*21] Shimada S., Golyanik V., Xu W., Pérez P., Theobalt C.: Neural monocular 3d human motion capture with physical awareness. *ACM Transactions on Graphics* (2021). 26

[SGXT20] Shimada S., Golyanik V., Xu W., Theobalt C.: Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics* (2020). 19

[SGY*21] Shen T., Gao J., Yin K., Liu M.-Y., Fidler S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems (NeurIPS)* (2021). 6

[SHM*18] Saito S., Hu L., Ma C., Luo L., Li H.: 3d hair synthesis using volumetric variational autoencoders. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* (2018). 22

[SHN*19] Saito S., Huang Z., Natsume R., Morishima S., Kanazawa A., Li H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision (ICCV)* (2019). 17

[SK87] Sirovich L., Kirby M.: Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A* (1987). 20

[Sla12] Slaughter W. S.: *The linearized theory of elasticity*. Springer Science & Business Media, 2012. 4

[SLF07] Salzmann M., Lepetit V., Fua P.: Deformable surface tracking ambiguities. In *Computer Vision and Pattern Recognition (CVPR)* (2007). 10

[SLNG20] Schwarz K., Liao Y., Niemeyer M., Geiger A.: Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)* (2020). 21, 22

[SMD*05] Stoyanov D., Mylonas G. P., Deligianni F., Darzi A., Yang G. Z.: Soft-tissue motion tracking and structure estimation for robotic assisted mis procedures. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2005). 14

[SNG*21] Shapovalov R., Novotny D., Graham B., Labatut P., Vedaldi A.: DensePose 3D: Lifting canonical surface maps of articulated objects to the third dimension. In *International Conference on Computer Vision (ICCV)* (2021). 16

[SOC22] Santesteban I., Otaduy M. A., Casas D.: SNUG: Self-Supervised Neural Dynamic Garments. *Computer Vision and Pattern Recognition (CVPR)* (2022). 19

[SPJG22] Song J., Patel M., Jasour A., Ghaffari M.: A closed-form uncertainty propagation in non-rigid structure from motion. *IEEE Robotics and Automation Letters* (2022). 14

[SSN*22] Schwarz K., Sauer A., Niemeyer M., Liao Y., Geiger A.: Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022). 21

[SSP07] Sumner R. W., Schmid J., Pauly M.: Embedded deformation for shape manipulation. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*. 2007. 7

[SSSJ20] Saito S., Simon T., Saragih J., Joo H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Computer Vision and Pattern Recognition (CVPR)* (2020). 1, 17

[STC21] Samsonov S., Tiampo K., Cassotto R.: Measuring the state and temporal evolution of glaciers in alaska and yukon using synthetic-aperture-radar-derived (sar-derived) 3d time series of glacier surface flow. *The Cryosphere* (2021). 26

[STG*20] Sidhu V., Tretschk E., Golyanik V., Agudo A., Theobalt C.: Neural dense non-rigid structure from motion with latent space constraints. In *European Conference on Computer Vision (ECCV)* (2020). 1, 13

[Sto12] Stoyanov D.: Stereoscopic scene flow for robotic assisted minimally invasive surgery. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2012). 13, 14

[STWW22] Skorokhodov I., Tulyakov S., Wang Y., Wonka P.: EpiGRAF: Rethinking training of 3d GANs. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022). 21

[SWY*22] Su Z., Wan W., Yu T., Liu L., Fang L., Wang W., Liu Y.: Mulaycap: Multi-layer human performance capture using a monocular video camera. *Transactions on Visualization and Computer Graphics* (2022). 18

[SY22] Shiohara K., Yamasaki T.: Detecting deepfakes with self-blended images. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 26

[SZW19] Sitzmann V., Zollhöfer M., Wetzstein G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems (NeurIPS)* (2019). 7, 8

[TBG*19] Tewari A., Bernard F., Garrido P., Bharaj G., Elgharib M., Seidel H.-P., Pérez P., Zöllhofer M., Theobalt C.: Fml: Face model learning from videos. In *Computer Vision and Pattern Recognition (CVPR)* (2019). 22

[TBGS16] Taetz B., Bleser G., Golyanik V., Stricker D.: Occlusion-aware video registration for highly non-rigid objects. In *Winter Conference on Applications of Computer Vision (WACV)* (2016). 14

[TBP*22] Tewari A., B R M., Pan X., Fried O., Agrawala M., Theobalt C.: Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 20, 21, 22

[TD20] Teed Z., Deng J.: Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision (ECCV)* (2020). 14

[TFT*20] Tewari A., Fried O., Thies J., Sitzmann V., Lombardi S., Sunkavalli K., Martin-Brualla R., Simon T., Saragih J., Niessner M., Pandey R., Fanello S., Wetzstein G., Zhu J.-Y., Theobalt C., Agrawala M., Shechtman E., Goldman D. B., Zollhöfer M.: State of the art on neural rendering. *Computer Graphics Forum (Eurographics State of the Art Reports)* (2020). 3

[THB08] Torresani L., Hertzmann A., Bregler C.: Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2008). 12, 13

[TK92] Tomasi C., Kanade T.: Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision (IJCV)* (1992). 14

[TKCM16] Tulsiani S., Kar A., Carreira J., Malik J.: Learning category-specific deformable 3d models for object reconstruction. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2016). 16

[TKG20] Tulsiani S., Kulkarni N., Gupta A.: Implicit mesh reconstruction from unannotated image collections. *arXiv preprint arXiv:2007.08504* (2020). 16

[TP91] Turk M., Pentland A.: Eigenfaces for recognition. *Journal of Cognitive Neuroscience* (1991). 20

[TPBF87] Terzopoulos D., Platt J., Barr A., Fleischer K.: Elastically deformable models. In *Proc. Conference on Computer Graphics and Interactive Techniques* (1987). 25

[TTG*21] Tretschk E., Tewari A., Golyanik V., Zollhöfer M., Lassner C., Theobalt C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *International Conference on Computer Vision (ICCV)* (2021). 2, 7, 14, 15

[TTM*22] TEWARI A., THIES J., MILDENHALL B., SRINIVASAN P., TRETSCHK E., YIFAN W., LASSNER C., SITZMANN V., MARTIN-BRUALLA R., LOMBARDI S., SIMON T., THEOBALT C., NIESSNER M., BARRON J. T., WETZSTEIN G., ZOLLHÖFER M., GOLYANIK V.: Advances in Neural Rendering. *Computer Graphics Forum (Eurographics State of the Art Reports)* (2022). 3, 14, 25, 26

[TTZ*20] TRETSCHK E., TEWARI A., ZOLLHÖFER M., GOLYANIK V., THEOBALT C.: Demea: Deep mesh autoencoders for non-rigidly deforming objects. In *European Conference on Computer Vision (ECCV)* (2020). 11

[TZG*18] TEWARI A., ZOLLHÖFER M., GARRIDO P., BERNARD F., KIM H., PÉREZ P., THEOBALT C.: Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Computer Vision and Pattern Recognition (CVPR)* (2018). 25

[TZK*17] TEWARI A., ZOLLÖFER M., KIM H., GARRIDO P., BERNARD F., PEREZ P., CHRISTIAN T.: MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *International Conference on Computer Vision (ICCV)* (2017). 8

[TZLW22] TIAN Y., ZHANG H., LIU Y., WANG L.: Recovering 3d human mesh from monocular images: A survey. *arXiv e-prints* (2022). 3

[VA13] VICENTE S., AGAPITO L.: Balloon shapes: Reconstructing and deforming objects with volume from images. In *International Conference on 3D Vision (3DV)* (2013). 23

[VBPP05] VLASIC D., BRAND M., PFISTER H., POPOVIĆ J.: Face transfer with multilinear models. *ACM Transactions on Graphics* (2005). 13, 14

[VCR*18] VAROL G., CEYLAN D., RUSSELL B., YANG J., YUMER E., LAPTEV I., SCHMID C.: BodyNet: Volumetric inference of 3D human body shapes. In *European Conference on Computer Vision (ECCV)* (2018). 17

[VGS16] VLADISLAV GOLYANIK A. S. M., STRICKER D.: Nrsfm-flow: Recovering non-rigid scene flow from monocular image sequences. In *British Machine Vision Conference (BMVC)* (2016). 14

[VSFU12] VAROL A., SALZMANN M., FUA P., URTASUN R.: A constrained latent variable model. In *Computer Vision and Pattern Recognition (CVPR)* (2012). 12, 14

[VWB*12] VALGAERTS L., WU C., BRUHN A., SEIDEL H.-P., THEOBALT C.: Lightweight binocular facial performance capture under uncontrolled lighting. *ACM Transactions on Graphics* (2012). 12, 14

[WBM*16] WOOD E., BALTRUŠAITIS T., MORENCY L.-P., ROBINSON P., BULLING A.: A 3d morphable eye region model for gaze estimation. In *European Conference on Computer Vision (ECCV)* (2016). 22

[WBW*11] WAH C., BRANSON S., WELINDER P., PERONA P., BELONGIE S.: *Caltech-UCSD Birds-200-2011 (CUB-200-2011)*. Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011. 2, 16

[WCF07] WHITE R., CRANE K., FORSYTH D. A.: Capturing and animating occluded cloth. *ACM Transactions on Graphics* (2007). 14

[WCL*22] WU Y., CHEN Z., LIU S., REN Z., WANG S.: CASA: Category-agnostic skeletal animal reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022). 24

[WCS*22] 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 *Computer Vision and Pattern Recognition (CVPR)* (2022). 17

[WCZ*22] WANG D., CHANDRAN P., ZOSS G., BRADLEY D., GOTARDO P.: Morf: Morphable radiance fields for multiview neural head modeling. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* (2022). 20, 21, 22

[WJ97] WADE M., JONES G.: The role of vision and spatial orientation in the maintenance of posture. *Physical Therapy* (1997). 1

[WJRV21] WU S., JAKAB T., RUPPRECHT C., VEDALDI A.: DOVE: Learning deformable 3d objects by watching videos. *arXiv preprint arXiv:2107.10844* (2021). 16, 17

[WKDB21] WANG Y., KOLOTOUROS N., DANIILIDIS K., BADGER M.: Birds of a feather: Capturing avian shape models from images. In *Computer Vision and Pattern Recognition (CVPR)* (2021). 1, 24

[WL21] WANG C., LUCEY S.: Paul: Procrustean autoencoder for unsupervised lifting. In *Computer Vision and Pattern Recognition (CVPR)* (2021). 13, 14

[WLL*21] WANG P., LIU L., LIU Y., THEOBALT C., KOMURA T., WANG W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)* (2021). 6, 8, 15

[WLM*22] WANG J., LUVIZON D., MUELLER F., BERNARD F., KORTYLEWSKI A., CASAS D., THEOBALT C.: Handflow: Quantifying view-dependent 3d ambiguity in two-hand reconstruction with normalizing flow. In *International Symposium on Vision, Modeling, and Visualization (VMV)* (2022). 23, 24

[WLPL22] WANG C., LI X., PONTES J. K., LUCEY S.: Neural prior for trajectory estimation. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 13, 14

[WMB*20] WANG J., MUELLER F., BERNARD F., SORLI S., SOTNYCHENKO O., QIAN N., OTADUY M. A., CASAS D., THEOBALT C.: Rgb2hands: Real-time tracking of 3d hand interactions from monocular rgb video. *ACM Transactions on Graphics* (2020). 23, 24

[WOR11] WANG H., O'BRIEN J. F., RAMAMOORTHI R.: Data-driven elastic models for cloth: modeling and measurement. *ACM Transactions on Graphics* (2011). 11

[WRV20] WU S., RUPPRECHT C., VEDALDI A.: Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Computer Vision and Pattern Recognition (CVPR)* (2020). 9, 16, 17, 20

[WTP03] WEMPNER G., TALASLIDIS D., PETROLITO J.: Mechanics of solids and shells: theories and approximations. *Appl. Mech. Rev.* (2003). 5

[WZA*22] WUU C.-H., ZHENG N., ARDISSON S., BALI R., BELKO D., BROCKMEYER E., EVANS L., GODISART T., HA H., HYPES A., KOSKA T., KRENN S., LOMBARDI S., LUO X., MCPHAIL K., MILLERSCHOEN L., PERDOCH M., PITTS M., RICHARD A., SARAGIH J., SARAGIH J., SHIRATORI T., SIMON T., STEWART M., TRIMBLE A., WENG X., WHITEWOLF D., WU C., YU S.-I., SHEIKH Y.: Multiface: A dataset for neural face rendering. In *arXiv* (2022). 22

[XAS21] XU H., ALLDIECK T., SMINCHISESCU C.: H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In *Advances in Neural Information Processing Systems (NeurIPS)* (2021). 18

[XBZ*20] XU H., BAZAVAN E. G., ZANFIR A., FREEMAN W. T., SUKTHANKAR R., SMINCHISESCU C.: Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Computer Vision and Pattern Recognition (CVPR)* (2020). 18

[XCZ*18] XU W., CHATTERJEE A., ZOLLHÖFER M., RHODIN H., MEHTA D., SEIDEL H.-P., THEOBALT C.: Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics* (2018). 18

[XHKK21] XIAN W., HUANG J.-B., KOPF J., KIM C.: Space-time neural irradiance fields for free-viewpoint video. In *Computer Vision and Pattern Recognition (CVPR)* (2021). 15

[XPWH20] XIANG D., PRADA F., WU C., HODGINS J. K.: Monoclothcap: Towards temporally coherent clothing capture from monocular RGB video. In *International Conference on 3D Vision (3DV)* (2020). 18

[XPY*22] Xu Y., Peng S., Yang C., Shen Y., Zhou B.: 3d-aware image synthesis via learning structural and textural representations. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 21, 22

[XTS*22] Xie Y., Takikawa T., Saito S., Litany O., Yan S., Khan N., Tombari F., Tompkin J., Sitzmann V., Sridhar S.: Neural fields in visual computing and beyond. *Computer Graphics Forum (Eurographics State of the Art Reports)* (2022). 3

[XX22] Xia W., Xue J.-H.: A survey on 3d-aware image synthesis. *arXiv e-prints* (2022). 3

[XXG*20] Xu L., Xu W., Golyanik V., Habermann M., Fang L., Theobalt C.: Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In *Computer Vision and Pattern Recognition (CVPR)* (2020). 25

[XYDT22] Xiang J., Yang J., Deng Y., Tong X.: Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. In *arXiv* (2022). 20, 21, 22

[XYTB22] Xiu Y., Yang J., Tzionas D., Black M. J.: ICON: Implicit Clothed humans Obtained from Normals. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 18

[YCLK20] Yang J., Chang H. J., Lee S., Kwak N.: Seqhand:rgb-sequence-based 3d hand pose and shape estimation. In *European Conference on Computer Vision (ECCV)* (2020). 23

[YGT22] Ye Y., Gupta A., Tulsiani S.: What's in your hands? 3d reconstruction of generic objects in hands. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 24

[YHL*22] Yao C.-H., Hung W.-C., Li Y., Rubinstein M., Yang M.-H., Jampani V.: Lassie: Learning articulated shapes from sparse image ensemble via 3d part discovery. *Advances in Neural Information Processing Systems (NeurIPS)* (2022). 4, 7, 8, 16

[YKG*20] Yoon J. S., Kim K., Gallo O., Park H. S., Kautz J.: Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *Computer Vision and Pattern Recognition (CVPR)* (2020). 15, 25

[YPA*18] Yang S., Pan Z., Amert T., Wang K., Yu L., Berg T., Lin M. C.: Physics-inspired garment recovery from a single-view image. *ACM Transactions on Graphics* (2018). 18

[YRCA15] Yu R., Russell C., Campbell N. D. F., Agapito L.: Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video. In *International Conference on Computer Vision (ICCV)* (2015). 11, 12, 23, 25

[YSJ*21a] Yang G., Sun D., Jampani V., Vlasic D., Cole F., Chang H., Ramanan D., Freeman W. T., Liu C.: Lasr: Learning articulated shape reconstruction from a monocular video. In *Computer Vision and Pattern Recognition (CVPR)* (2021). 15

[YSJ*21b] Yang G., Sun D., Jampani V., Vlasic D., Cole F., Liu C., Ramanan D.: Viser: Video-specific surface embeddings for articulated 3d shape reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)* (2021). 1, 8, 16

[YSL*22] Yuan Y.-J., Sun Y.-T., Lai Y.-K., Ma Y., Jia R., Gao L.: Nerf-editing: Geometry editing of neural radiance fields. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 26

[YSZZ19] Yang L., Shi Z., Zheng Y., Zhou K.: Dynamic hair modeling from monocular videos using deep neural networks. *ACM Transactions on Graphics* (2019). 22

[YVN*22] Yang G., Vo M., Natalia N., Ramanan D., Andrea V., Hanbyul J.: Banmo: Building animatable 3d neural models from many casual videos. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 7, 8, 15, 16

[YWS*06] Yin L., Wei X., Sun Y., Wang J., Rosato M.: A 3d facial expression database for facial behavior research. In *Proc. International Conference on Automatic Face and Gesture Recognition* (2006). 13

[YZH*22] Yi X., Zhou Y., Habermann M., Shimada S., Golyanik V., Theobalt C., Xu F.: Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 19

[YZW*20] Yang H., Zhu H., Wang Y., Huang M., Shen Q., Yang R., Cao X.: Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Computer Vision and Pattern Recognition (CVPR)* (2020). 22

[ZAB*22] Zheng Y., Abrevaya V. F., Bühler M. C., Chen X., Black M. J., Hilliges O.: I M Avatar: Implicit morphable head avatars from videos. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 20, 22

[ZBL*19] Zhou Y., Barnes C., Lu J., Yang J., Li H.: On the continuity of rotation representations in neural networks. In *Computer Vision and Pattern Recognition (CVPR)* (2019). 7

[ZBT22] Zielonka W., Bolkart T., Thies J.: Towards metrical reconstruction of human faces. In *European Conference on Computer Vision (ECCV)* (2022). 20

[ZCY*19] Zimmermann C., Ceylan D., Yang J., Russell B., Argus M., Brox T.: Freihand: Dataset for markerless capture of hand pose and shape from single rgb images. In *International Conference on Computer Vision (ICCV)* (2019). 23, 24

[ZDY*21] Zeng H., Dai Y., Yu X., Wang X., Yang Y.: Pr-rrn: Pairwise-regularized residual-recursive networks for non-rigid structure-from-motion. In *International Conference on Computer Vision (ICCV)* (2021). 14

[ZEJ*16] Zolfaghari R., Epain N., Jin C., Glaunés J., Tew A.: Generating a morphable model of ears. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016). 22

[ZGZ*21] Zou S., Guo C., Zuo X., Wang S., Wang P., Hu X., Chen S., Gong M., Cheng L.: Eventhpe: Event-based 3d human pose and shape estimation. In *International Conference on Computer Vision (ICCV)* (2021). 25

[ZHDLTL14] Zhu Y., Huang D., De La Torre F., Lucey S.: Complex non-rigid motion 3d reconstruction by union of subspaces. In *Computer Vision and Pattern Recognition (CVPR)* (2014). 13

[ZHX*18] Zhou Y., Hu L., Xing J., Chen W., Kung H.-W., Tong X., Li H.: Hairnet: Single-view hair reconstruction using convolutional neural networks. In *European Conference on Computer Vision (ECCV)* (2018). 22

[ZHX*20] Zhou Y., Habermann M., Xu W., Habibie I., Theobalt C., Xu F.: Monocular real-time hand shape and motion capture using multi-modal data. In *Computer Vision and Pattern Recognition (CVPR)* (2020). 23

[ZIE*18] Zhang R., Isola P., Efros A. A., Shechtman E., Wang O.: The unreasonable effectiveness of deep features as a perceptual metric. In *Computer Vision and Pattern Recognition (CVPR)* (2018). 8

[ZKB18] Zuffi S., Kanazawa A., Black M. J.: Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In *Computer Vision and Pattern Recognition (CVPR)* (2018). 24

[ZKBWB19] Zuffi S., Kanazawa A., Berger-Wolf T., Black M. J.: Three-D safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In *International Conference on Computer Vision (ICCV)* (2019). 24

[ZKJB17] Zuffi S., Kanazawa A., Jacobs D., Black M. J.: 3D menagerie: Modeling the 3D shape and pose of animals. In *Computer Vision and Pattern Recognition (CVPR)* (2017). 7, 24

[ZLM*19] Zhang X., Li Q., Mo H., Zhang W., Zheng W.: End-to-end hand mesh recovery from a monocular rgb image. In *International Conference on Computer Vision (ICCV)* (2019). 23

[ZPL*22] Zhu Z., Peng S., Larsson V., Xu W., Bao H., Cui Z., Oswald M. R., Pollefeys M.: Nice-slam: Neural implicit scalable encoding for slam. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 25

[ZRSK20] Zhang K., Riegler G., Snavely N., Koltun V.: Nerf++: Analyzing and improving neural radiance fields. *ArXiv* (2020). 25

[ZSD*21] Zhang X., Srinivasan P. P., Deng B., Debevec P., Freeman W. T., Barron J. T.: Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics* (2021). 5

[ZSG*18] Zollhöfer M., Stotko P., Görlitz A., Theobalt C., Niessner M., Klein R., Kolb A.: State of the art on 3d reconstruction with rgb-d cameras. *Computer Graphics Forum (Eurographics State of the Art Reports)* (2018). 2

[ZTG*18] Zollhöfer M., Thies J., Garrido P., Bradley D., Beeler T., Pérez P., Stamminger M., Niessner M., Theobalt C.: State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer Graphics Forum (Eurographics State of the Art Reports)* (2018). 3, 20

[ZTTT00] Zienkiewicz O. C., Taylor R. L., Taylor R. L., Taylor R. L.: *The finite element method: solid mechanics*, vol. 2. Butterworth-heinemann, 2000. 5

[ZWD*21] Zhang B., Wang Y., Deng X., Zhang Y., Tan P., Ma C., Wang H.: Interacting two-hand 3d pose and shape reconstruction from single color image. In *International Conference on Computer Vision (ICCV)* (2021). 24

[ZWZ*22] Zhu H., Wu W., Zhu W., Jiang L., Tang S., Zhang L., Liu Z., Loy C. C.: CelebV-HQ: A large-scale video facial attributes dataset. In *European Conference on Computer Vision (ECCV)* (2022). 22, 23

[ZXLK21] Zhai M., Xiang X., Lv N., Kong X.: Optical flow and scene flow estimation: A survey. *Pattern Recognition* (2021). 8

[ZYHC22] Zheng M., Yang H., Huang D., Chen L.: Imface: A nonlinear 3d morphable face model with implicit neural representations. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 20

[ZYLD21] Zheng Z., Yu T., Liu Y., Dai Q.: Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2021). 18

[ZYW*19] Zheng Z., Yu T., Wei Y., Dai Q., Liu Y.: Deephuman: 3d human reconstruction from a single image. In *International Conference on Computer Vision (ICCV)* (2019). 17

[ZZL*22] Zhao H., Zhang J., Lai Y.-K., Zheng Z., Xie Y., Liu Y., Li K.: High-fidelity human avatars from a single rgb camera. In *Computer Vision and Pattern Recognition (CVPR)* (2022). 18

[ZZSC22] Zhuang Y., Zhu H., Sun X., Cao X.: Mofanerf: Morphable facial neural radiance field. In *European Conference on Computer Vision (ECCV)* (2022). 20, 21, 22