High-Fidelity 4D Cloth Capture Pipeline with a Two-Level Pattern

ZIHENG LIU, University of Utah, USA
ANKA CHEN, NVIDIA, USA
SHU CHEN, University of Utah, USA
YIN YANG, University of Utah, USA
CEM YUKSEL, University of Utah, USA
JENNY HAN LIN, University of Utah, USA

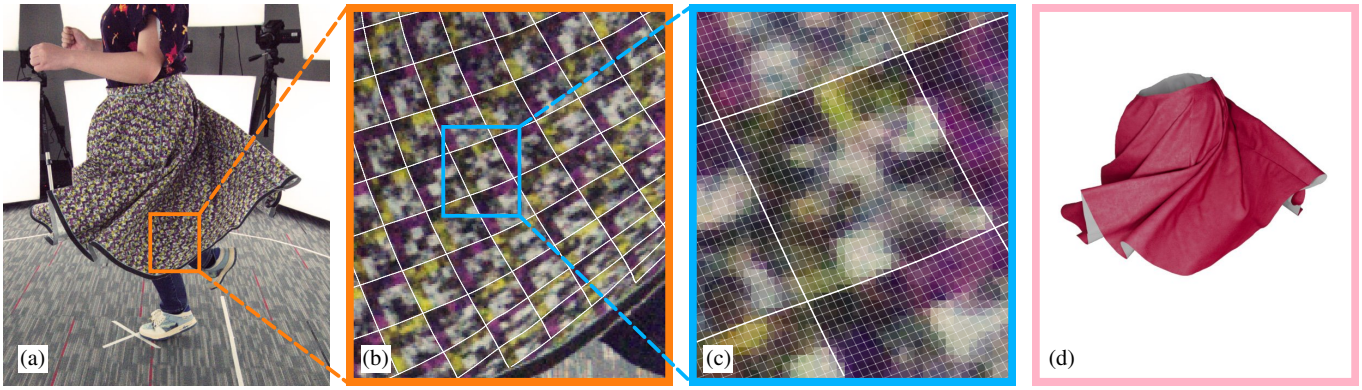


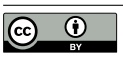
Fig. 1. *Our 4D cloth capture pipeline can reconstruct fast, complex motions such as (a) this spinning circle skirt. This is achieved using (b) sparse markers that provide coarse localization and registration using (c) dense noise patterns where crosspoints are localized with 1mm resolution. Our novel post-processing method inpaints occluded regions to produce (d) intersection-free meshes that reproduce fine wrinkles and folds.*

Capturing cloth motion with high fidelity is challenging due to fine-scale wrinkles, large deformation, and frequent self-occlusion. We present a 4D (spatio-temporal) cloth capture system that achieves 1 mm spatial resolution using only 16 RGB cameras. Our approach uses a two-level marker pattern: sparse, colored L-shaped markers provide robust detection and orientation, while dense noise patterns within each marker enable both marker identification and precise keypoint localization. By unwarping detected markers to a canonical frame, we factor out perspective distortion and most of cloth deformation, allowing the localizer to achieve sub-pixel accuracy. The localized keypoints are triangulated across views to form an incomplete point cloud. A physics-based optimization then deforms a template mesh to match the captured geometry while maintaining penetration-free constraints and physical plausibility for occluded regions. Our method produces temporally coherent sequences that faithfully capture fine wrinkles and folds even during complex motions with self-contact.

CCS Concepts: • **Computing methodologies** → **Motion capture**.

ACM Reference Format:
Ziheng Liu, Anka Chen, Shu Chen, Yin Yang, Cem Yuksel, and Jenny Han Lin. 2026. High-Fidelity 4D Cloth Capture Pipeline with a Two-Level Pattern. *ACM Trans. Graph.* 45, 4 (July 2026), 12 pages. https://doi.org/10.1145/3811305

Authors' Contact Information: Ziheng Liu, University of Utah, Salt Lake City, UT, USA, lzh2199@gmail.com; Anka Chen, NVIDIA, Kirkland, USA, ankachan92@gmail.com; Shu Chen, University of Utah, USA, samuel233qaq@gmail.com; Yin Yang, University of Utah, USA, yangzzzy@gmail.com; Cem Yuksel, University of Utah, USA, cem@cemyuksel.com; Jenny Han Lin, University of Utah, USA, jenny.h.lin@utah.edu.



ACM 1557-7368/2026/7-ART
https://doi.org/10.1145/3811305

1 Introduction

As data-driven methods become more prevalent in computer graphics, they accentuate the importance of capturing real-world data. The resulting data serve as valuable assets for a wide range of applications, including character animation [Feng et al. 2010; Wang et al. 2010], virtual try-on [Chong et al. 2021; Islam et al. 2024; Patel et al. 2020; Santesteban et al. 2019], and digital fashion [Baek et al. 2022]. Unfortunately, getting sufficiently detailed capture of cloth dynamics has remained an open problem. Cloth has many fine-scale details, such as wrinkles and folds, and in dynamic settings it frequently comes into self-contact and generates self-occlusions. Furthermore, in garments, the cloth is also interacting with the human body and has complicated dynamics based on human motions and garment construction. While there have been improvements in capturing high-fidelity cloth motion, many methods require expensive capture setups with hundreds of cameras [Halimi et al. 2022a]. Thus, the lack of fine-grained and temporally coherent 4D cloth data continues to hinder progress in data-driven research across related domains.

Among existing approaches, marker-based cloth capture offers the highest resolution and accuracy. Such methods use special visual markers printed on the cloth to provide reliable features for detection and localization. While the dependence on markers makes these methods less generalizable than markerless methods, which rely on universal but sparse garment features, the addition of clear, dense markers vastly simplifies the problem of establishing consistent correspondences across views and over time, resulting in high geometric accuracy in the final reconstruction. Thus despite the increased cost required to construct specialized capture garments,

marker-based cloth capture is ideal for applications that require detailed, physically accurate cloth motion data.

However, existing marker-based methods face an additional hurdle that prevents wider adoption. High-resolution captures require small, densely distributed markers that can be detected and identified even as the cloth moves and deforms. While marker identification can be improved by using numerous cameras with varying viewpoints and focal lengths to capture cloth details at different scales [Halimi et al. 2022b], introducing more cameras quickly increases the cost and complexity of the system. Thus, methods with fewer cameras have focused on either static, close-up reconstructions [White et al. 2007] or dynamic capture with significantly lower resolution [Chen et al. 2021; Scholz et al. 2005; White et al. 2007]. These limitations underscore the need for a robust cloth capture solution that can achieve high-resolution reconstructions across diverse garment types and motion patterns, while remaining practical in terms of hardware and setup complexity.

Furthermore, missing regions are inevitable even in capture settings with many cameras. These gaps typically arise from occlusion, where certain surface areas are visible to fewer than two cameras. This is a particularly severe issue for clothing due to its complex, self-occluding geometry. When inpainting these missing regions, the geometry must not only be physically plausible, it must be spatially and temporally coherent with the captured geometry. Furthermore, the geometry should ideally be intersection-free, as otherwise non-physical artifacts will compromise both visual quality and downstream tasks. This is an issue for both the inpainted regions and the reconstructed geometry, which may have small intersections due to noise in the capture. Solving these issues is particularly difficult due to the high resolution of the captured garment and the need for temporal coherence.

In this paper, we propose a new end-to-end pipeline for capturing and reconstructing high-resolution garment motion. Our system begins with a two-level pattern design and a corresponding localization pipeline for high-fidelity cloth capture from multiview RGB images. The first level features a sparse checkerboard layout of L-shaped markers distributed across the garment. The second level overlays robust noise patterns within each marker, enabling effective keypoint localization across a range of resolutions. This design supports a two-stage localization process: we trained one network to detect and localize the sparse global markers, which are then unwarped to a canonical square. Then, we trained a classifier to identify each coarse marker and a keypoint localizer to place the dense keypoints (called *crosspoints*) within each canonical marker. This hierarchical formulation reduces the problem complexity at each stage and normalizes out perspective transformations, which neural networks often struggle to handle. This improves robustness, efficiency, and accuracy under varying capture distances, thus allowing good results with fewer cameras. In addition, all three networks were trained on synthetic data and can generalize across different garments, further improving the pipeline's applicability to diverse capture scenarios.

Once the markers have been localized and triangulated to generate the captured 3D positions, we use a novel physics-based optimization to produce a faithful, coherent motion sequence guaranteed to be free of self-penetrations. We do this by formulating energy terms that move captured vertices towards the positions from the triangulation, while occluded vertices are given energy terms that encourage them to resemble adjacent frames. Offset Geometric Contact (OGC) [Chen et al. 2025] is used to safeguard our optimization so the entire optimization is free of self-intersection. As a result, captured regions stay faithful to the triangulation result while potential self-intersections caused by capture noise are resolved. Meanwhile, inpainted regions are given reasonable cloth positions that are coherent with the captured part while also being temporally smooth and penetration-free.

Table 1. *Pipeline Comparison. Resolution describes real-world mm between reconstructed points, while drift rate uses optical flow to calculate temporal incoherence between frames in the UV space.*

	Ours	[Chen et al. 2021]	[Halimi et al. 2022b]
# Cameras	**16**	**16**	211
Resolution [mm]	**1.06**	30	2.67
Drift rate [mm/sec]	**0.75**	N/A	1.50
Penetration-free	**Yes**	No	No

With this approach, we achieve an unprecedented real-world spatial resolution of 1.06 mm with only 16 synchronized RGB cameras. Our pipeline guarantees a penetration-free motion sequence while achieving 2× better temporal coherence than prior work, as shown in Table 1. Code and trained models are available at https://github.com/Utah-Graphics-Lab/clothcap.

2 Related Work

Prior cloth capture techniques can be broadly categorized as either data-driven or photogrammetry-based approaches. Data-driven methods can estimate garment geometry from single views [Daněřek et al. 2017] or even images in the wild [Moon et al. 2022], and may additionally infer cloth texture [Saito et al. 2019, 2020], sewing patterns [Liu et al. 2023], or alternative geometry representations such as Gaussian Splatting [Hu et al. 2024]. While some methods incorporate physical or depth cues to improve reconstruction quality [Chen et al. 2015; Popa et al. 2009; Yang et al. 2018], data-driven approaches generally prioritize visual plausibility over geometric fidelity and depend on large-scale training data. We consider them complementary to our pipeline, which can provide high-quality motion examples to enhance data-driven methods.

Photogrammetry-based methods, on which we focus in this section, employ multi-view or depth cameras to reconstruct 3D geometry directly from visual data. By establishing a correspondence between image data and spatial points, 3D positions can be precisely measured. While these methods require complex, carefully calibrated capture setups, they produce high-quality results.

2.1 Markerless methods

Early methods for 3D cloth reconstruction focused on leveraging features inherent to the garment to establish correspondence. Bradley et al. [2008] employed stereo matching to reconstruct 3D keypoints from image pairs, generating a point cloud from which a surface mesh was recovered via surface reconstruction techniques. de Aguiar et al. [2008] reconstructed a clothed human mesh by aligning a laser-scanned template mesh to input video frames. Vlasic

et al. [2008] fit the skeleton of a skinned template mesh to the input images and subsequently refined the mesh shape to better align with image silhouettes. More recent approaches have used shaped priors based on either the underlying human body [Pons-Moll et al. 2017] or captures of reference garments [Guo et al. 2023] to infer correspondence.

While these methods can be used to reconstruct any garment, markerless methods generally suffer from lower spatial resolution and reduced accuracy compared to marker-based counterparts. Furthermore, the need to maintain temporal coherence from frame to frame complicates the reconstruction and can produce temporal artifacts. These limitations have motivated continued development of marker-based methods, which trade the flexibility of capturing arbitrary garments for substantially higher accuracy by using specially printed patterns that provide reliable spatial and temporal correspondences.

2.2 Marker-based methods

Scholz et al. [2005] introduced a color-coded pattern composed of regular grids, where each grid cell was assigned a unique color. Building on this idea, subsequent works [Biasi et al. 2015; Halimi et al. 2022b] enhanced the robustness of the pattern design by incorporating constraints such as uniqueness and rotation invariance, and by increasing the spatial resolution of the encoded markers. White et al. [2007] proposed the use of tessellated triangles filled with random colors, along with an iterative algorithm to establish reliable inter-view correspondences. Chen et al. [2021] augmented a checkerboard-patterned suit with two human-readable characters per square to support keypoint annotation and recognition. More recently, Liang et al. [2024] introduced a design that integrates fiducial markers into a Pied-de-poule (houndstooth) pattern, allowing the garment to remain aesthetically pleasing while still being trackable.

Our work builds on marker-based approaches but introduces a two-level hierarchical pattern that decouples coarse localization from fine keypoint detection, enabling higher-resolution 4D capture with a practical camera setup. Furthermore, our physics-based post processing step directly prevents self-intersection in the reconstructed meshes. This is opposed to prior methods, which either provide some heuristic collision resolution for inpainted regions only [Halimi et al. 2022b], or perform reconstruction on relatively form-fitting garments, where occluded regions are less likely to have complex self-contact [Biasi et al. 2015; Chen et al. 2021; White et al. 2007]

3 Capturing with Two-Level Pattern

We propose a two-level pattern design for cloth capture, paired with a two-stage localization pipeline that detects a dense set of keypoints from RGB images. Leveraging multi-view observations, we triangulate the 3D positions of garment vertices and apply standard Laplacian smoothing to remove noise.

3.1 Pattern Design

The sparse-level pattern is designed to provide unique *localization points* used to unwarp the captured image for each marker to a canonical square. We arrange L-shaped markers, colored in either yellow or magenta, in a checkerboard layout across the garment surface, as illustrated in Fig. 2. The alternating colors help define boundaries between adjacent markers. The four corners of each marker's bounding region serve as unique localization points, while the L-shape's orientation encodes rotational information.

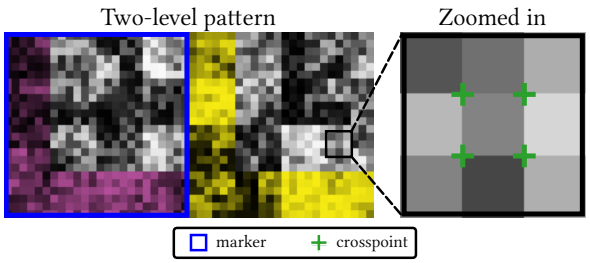
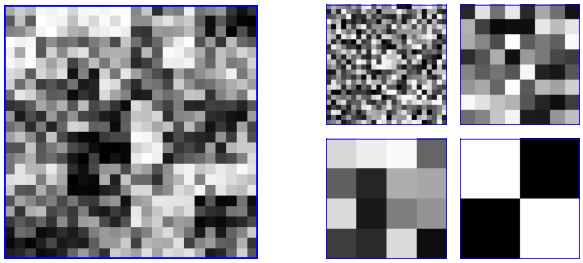


Fig. 2. *Two-level pattern design showing L-shaped markers (sparse level) with embedded crosspoint grid (dense level).*

The dense-level pattern serves two purposes: it provides dense localization points for improved capture resolution and uniquely encodes the identity of each marker. Inspired by the calibration board design in Li et al. [2013], we generate the dense pattern by superimposing multiple layers of Gaussian noise at different spatial frequencies, as shown in Fig. 3. When the cloth is captured at an optimal distance and in focus, the finest level of the pattern can be reliably extracted. Under lower capture quality, where high-frequency components become indistinguishable due to blur, the finer noise components tend to cancel out, allowing the coarser layers to stand out. This multi-frequency design improves robustness to motion and focus blur, thereby enhancing the reliability of crosspoint detection under real-world conditions and making the pattern particularly well-suited for high-fidelity motion capture.



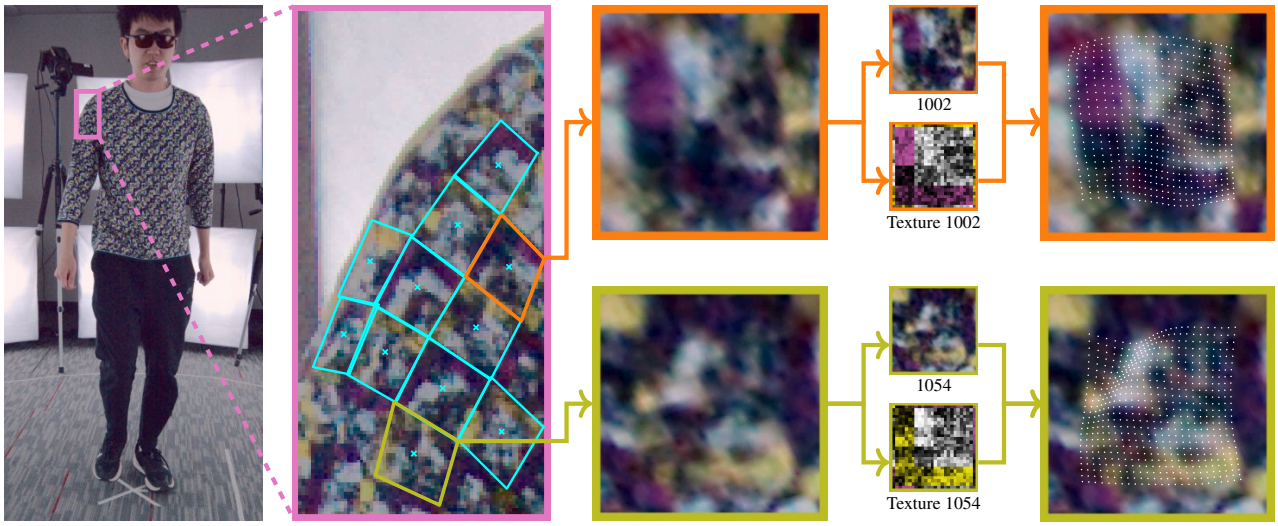
(a) *Dense Noise Pattern* **(b)** *Layers of Noise*

Fig. 3. *The dense pattern is generated by noise at different resolutions.*

The intersections of adjacent finest-level noise elements are referred to as *crosspoints*, as illustrated in Fig. 2. Each marker contains a 25-by-25 grid of crosspoints. The spacing between adjacent crosspoints is 6 pixels, corresponding to a physical resolution of about 1 mm when printed at 150 pixels per inch.

3.2 Pattern Localization Pipeline

We adopt a hierarchical two-stage localization approach in our pipeline. In the first stage, we detect and localize the four corner points of all markers in the image. In the second stage, the input image of each marker is unwarped and cropped based on the detected



(a) *Original Photo* **(b)** *Marker Localizer* **(c)** *Homography* **(d)** *Recognizer* **(e)** *Crosspoint Localizer*

Fig. 4. *Our pattern recognition pipeline begins by using our (b) Marker Localizer to find coarse markers in the image space. A (c) homography transformation is applied so that (d) marker recognition and (e) crosspoint localization are performed on a canonical view.*

localization points. A crosspoint localization process is conducted within each detected marker to precisely localize the crosspoints. This hierarchical design effectively reduces the problem size at each stage by reducing the variation of the input, ensuring that the workload remains manageable and allowing both stages to operate with improved efficiency and accuracy. Our full pipeline comprises three neural networks, as illustrated in Fig. 4. Critically, all training data is synthetically generated, as manual annotation is infeasible given the density of feature points. Furthermore, the Marker Localizer and Crosspoint Localizer networks in our pipeline generalize across garments and only need to be trained once. Only the Marker Recognizer must be retrained, and this is only if the garment contains new, unseen markers. We discuss these points further in our evaluation (Sec. 5). Architectural details for each network are provided in Sec. 1 of the supplemental.

The first network, Marker Localizer (Fig. 4b), is responsible for the initial stage of localization point detection. We adapt a YOLO-like architecture for this task [Redmon et al. 2016]. The network divides the image into cells and predicts whether the center of a marker falls within each cell. For positive cells, it further regresses the coordinates of the marker's center as well as its four unique localization points.

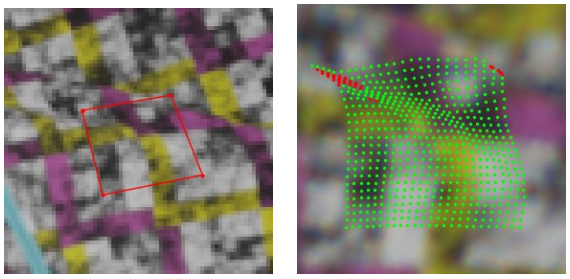
Before the second-stage localization, we apply a homography transformation to unwarp each detected marker into a standardized square (Fig. 4c). This transformation factors out all the perspective transformations and most of the non-rigid deformations such as stretching and shearing in the cloth image—factors that neural networks typically struggle to generalize across. As a result, it simplifies downstream processing by producing fixed-size input images for each marker with substantially reduced variance.

Following the unwarping, we introduce a marker recognition step to support subsequent crosspoint localization (Fig. 4d). The Marker Recognizer adopts a classic CNN+MLP architecture. Specifically, a ResNet backbone is used to extract features from the standardized marker image, and a multi-layer perceptron (MLP) then classifies these features into their corresponding marker IDs.

With the marker identity determined, we retrieve its reference texture and use it alongside the unwarped marker image as input to the Crosspoint Localizer (Fig. 4e). This formulation generalizes better across different marker appearances, improving both accuracy and efficiency, as the network does not need to memorize each marker internally. The network consists of two CNN branches that independently extract features from the warped image crop and the corresponding reference texture. These feature maps are then concatenated and fed into a U-Net architecture, which performs hierarchical encoding and decoding to produce dense localization outputs. For each of the 25×25 crosspoints, the network outputs three values: a visibility flag indicating whether the crosspoint is visible in the image, and, if visible, the (x, y) coordinates of the crosspoint within the unwarped marker image. Finally, by applying the inverse homography transformation, we project these coordinates back to the original image space.

Note that the homography approximation assumes each marker region is approximately planar, and crosspoint localization does degrade as the marker geometry deviates from this assumption. However, our pipeline is still capable of reconstructing moderate wrinkles smaller than the size of a single marker. This is because the Crosspoint Localizer is trained on synthetic data covering a wide range of deformations, including heavily non-planar cases. Crosspoints within the marker that can be confidently identified can be used for reconstruction, while crosspoints that are occluded or otherwise unrecoverable are marked as invalid, excluding them from triangulation; the resulting gaps are completed by the physics-based post-processing. Fig. 5 illustrates this behavior on a synthetic example.

Additionally, we enable crosspoint localization for partially visible markers. Although the four localization points of these markers are not all observable, we can estimate a reasonable homography



(a) *Zoomed-in Marker* **(b)** *Homography Result*

Fig. 5. *(a) A zoomed-in view of a wrinkled marker; red quads indicate the corner points used for the homography transform. (b) The homography-rectified canonical view with crosspoints overlaid. Surface folds introduce geometric distortion in the rectified image; crosspoints that are occluded or in close proximity to their neighbors are flagged as invalid (shown in red).*

transformation using adjacent fully visible markers. The resulting unwarped image and its corresponding texture are then fed into the same crosspoint localization network as used for full markers. This strategy allows us to detect crosspoints in partially occluded markers, which commonly appear near garment boundaries.

3.3 Triangulation

We employ a standard triangulation technique to recover the 3D position of each crosspoint that is observed by at least two cameras. Each estimated 3D point is reprojected back into the image planes using the corresponding camera parameters. To refine the 3D position, we minimize the reprojection error across all relevant views—defined as the mean squared error between the observed and reprojected 2D coordinates.

To further improve accuracy, we apply additional filtering to remove outliers. For each point, outlier observations are removed using RANSAC, which identifies them based on their reprojection errors. Furthermore, crosspoints with a final reprojection error exceeding an empirically determined threshold are excluded from the reconstruction. Finally, we apply a standard Laplacian smoothing to remove spatial noise while preserving details in the captured data.

4 Physics-Based Post-Processing

The captured high-resolution garment geometry is useful for potential downstream applications such as simulation, garment retargeting, and dataset generation for learning-based methods. However, the unprocessed data is not ready to be used directly due to missing regions caused by self-occlusions and potentially intersecting geometry due to capture noise. To that end, we present a physics-based post-processing pipeline that completes missing regions and removes intersection artifacts while preserving data fidelity.

Our post-processing pipeline builds upon two existing physics-based simulation techniques: Vertex Block Descent (VBD) [Chen et al. 2024] and Offset Geometric Contact (OGC) [Chen et al. 2025]. VBD is a highly parallel GPU-based solver, while OGC provides a penetration-free contact handling framework particularly well-suited for cloth self-collision. These two methods are compatible and have been integrated into the open-source Newton Physics Engine [Newton Contributors 2025], which serves as the foundation of our pipeline.

We introduce several key modifications to adapt the simulator for our specific purpose. First, we use dynamic rest shapes to preserve the captured regions while completing missing parts and maintaining temporal coherence. Second, we adopt a multi-resolution approach to mitigate VBD's slow convergence when handling meshes with large degrees of freedom under large timesteps. Further implementation details are provided in Sec. 2 of the supplemental.

4.1 Penetration-Free Simulation

Given a penetration-free initial state, OGC ensures a simulation free of self-penetration. Thus, we can start with a penetration-free default garment shape and use OGC to progressively transform the mesh to match the captured data. This is done by adding a spring-like energy based on the distance between each reconstructed vertex's current position and its target position, i.e., the triangulated location from the capture. Since the template is derived directly from the 2D sewing pattern, constructing it in a penetration-free state is straightforward. We select an initial frame with a high reconstruction rate and a shape close to the template, which helps the optimization converge to a consistent starting pose. In our target use case of controlled capture environments, the subject can strategically perform such a pose, making this requirement easy to satisfy. Even without such a frame, a low reconstruction rate in the initial frame only affects convergence difficulty and inpainting accuracy; the result remains penetration-free regardless.

From there, we employ the cloth material and barrier contact models built into the Newton simulator. These include the Saint Venant-Kirchhoff (StVK) model [Irving et al. 2004] for membrane deformation and dihedral-angle bending [Grinspun et al. 2003], which together govern stretching, shearing, bending, and contact response. This initial frame can then be used to propagate the simulation forward and backward through the rest of the capture sequence. This formulation drives captured vertices toward their target locations while keeping the missing regions physically plausible, all while maintaining a penetration-free state throughout.

4.2 Smoothing with Dynamic Rest Shape

While cloth simulation typically uses a fixed rest shape, where deviations from this rest shape are penalized, doing so introduces undesirable artifacts. In the reconstructed regions, larger discrepancies with the captured data begin to emerge. This makes sense, as we are not actually trying to simulate the behavior of a known material, but instead match the behavior of our captured data. Thus, for captured regions, we use the dihedral angles computed from the captured data as the rest angles for bending and derive the deformation gradient from the captured state. In this way, the bending and membrane energies do not penalize the captured geometry, as the rest configuration already matches the observed shape.

Meanwhile, in the inpainted regions, using a fixed rest shape produces temporal jittering. This is likely due to the sim-to-real gap in the cloth model causing jumps at boundaries with captured regions. Thus, for missing regions, we use the geometry from the

previous frame as the rest shape. This encourages minimal deformation from the previous state, thereby enhancing temporal coherence across the sequence. Additionally, we further smooth the missing regions by applying extra rounds of simulation after the initial pass: using the average of both the previous and following frames as the rest shape, leveraging information from both temporal directions. We also set the spring energy to be significantly stronger than the bending and membrane energies, ensuring that the captured regions remain faithful to the observations.

Our smoothing technique is essential for generating high-quality cloth data. By carefully balancing different rest shape strategies for captured and missing regions, we preserve the fidelity of observed geometry while ensuring that completed regions remain spatially smooth, temporally coherent, and physically plausible.

4.3 Multi-Resolution Simulation

Although VBD excels at simulating scenes with a large number of vertices, it suffers from slow convergence or even failure when all those vertices are contained in a single object, particularly under large timesteps. This is because VBD works best at minimizing local error and struggles to reduce overall global error. However, we can adapt VBD to work in this scenario by using a multi-resolution approach. Solving first on a coarse mesh produces a global shape that is close to convergence. This allows later iterations to efficiently refine local details and converge to a final solution.

For each garment, we construct a hierarchy of meshes at different resolutions by progressively downsampling the full-resolution mesh. The coarsest level contains only thousands of vertices, which is well within VBD's capability even under a large timestep. After applying the aforementioned simulation and smoothing strategy at this coarsest level, we generate the next finer level through simple interpolation. This process is repeated at each subsequent resolution. Although the vertex count increases at finer levels, the global low-frequency errors have already been resolved, leaving only local high-frequency errors that VBD handles efficiently. By repeating this strategy until the finest level, we recover the full-resolution mesh using VBD.

5 Evaluation

We begin this section by describing the setup used to capture our real-world garment motion data as well as our synthetic data generation process. We then perform a qualitative and quantitative analyses of the quality of out pipeline's reconstructions before examing the performance of our component networks on synthetic data. Furthermore, we examine the importance of our two-level pattern by analyzing its robustness to blur and presenting ablation studies that use only one level of our pattern. Finally, we present computation times.

To evaluate our pipeline, we prepared four garments featuring our custom marker pattern: a long-sleeved T-shirt, a circle skirt, a full-body suit, and a pair of pants, referred to as the *shirt*, *skirt*, *suit*, and *pants*, respectively. The physical shirt and suit are made of knitted polyester-spandex fabric, which is elastic and offers high print fidelity, priced at 36 dollars per yard. The skirt and pants are made of woven cotton, a stiffer material that exhibits more plastic deformation and lower print quality, priced at 20 dollars per yard. The total cost of all four garments is less than 300 dollars. We printed our marker pattern onto the fabric, then cut and sewed the pieces to construct the garments. Despite differences in fabric properties and garment types, our pattern and pipeline remain effective across all cases, demonstrating strong robustness. We also constructed digital meshes of our four garments textured with our marker pattern. These digital garments were simulated on SMPL body meshes [Loper et al. 2015], animated with AMASS motions [Mahmood et al. 2019] using a physics-based cloth simulator [Lu et al. 2025], and rendered under varied lighting to generate training images with automatic ground-truth labels. Further details on garment design and synthetic data generation are provided in Sec. 3 of the supplemental.

We then captured four motion sequences, each performed by a subject wearing one of the garments, with one sequence featuring the subject wearing both the shirt and pants simultaneously to demonstrate our pipeline's ability to handle multiple garments. Our multi-camera setup consists of 16 standard RGB cameras arranged in a circle around the capture volume. For the skirt capture, all cameras are positioned at the same height, focusing on the lower body. For the other three garments, four of the 16 cameras are elevated above head level to better capture upper-body details. Each camera captures images at a resolution of 4000×2160 at 30 fps. Shutters are synchronized via genlock with negligible synchronization error. To ensure consistent and diffuse illumination, we place 30 softboxes around the capture volume. The bright lighting allows us to use a small aperture and a fast 1/2000 s shutter speed, resulting in sharp images even during fast motion. All cameras are calibrated using a standard checkerboard-based method. The motions include both standard movements (e.g., T-pose, walking) and dynamic, complex actions (e.g., pitching, martial arts, jumping, spinning), resulting in diverse garment deformations. Our method performs consistently well under all conditions, further validating its robustness.

5.1 Reconstruction Quality

Directly comparing 3D reconstruction methods is difficult due to variations between capture setups, motion sequences, and human actors. In this section, we provide some quantitative comparison against metrics reported by Halimi et al. [2022a] and Chen et al. [2021], as well as providing our own analysis on the quality of our reconstruction. Particularly challenging poses are shown in Fig. 9, where the high level of detail in the direct reconstruction as well as the quality of the inpainting can be compared. Recall that all our reconstructions are penetration-free even in the face of complex wrinkles and self-contact.

UV-Space Optical Flow. UV-space optical flow drift rate is a metric introduced by Halimi et al. [2022b] to provide reproducible evaluation of reconstruction accuracy invariant to specific multi-view camera configurations. It is calculated by projecting all captured images into UV space using the triangulated point cloud, producing a single merged texture. UV-space optical flow is then computed between consecutive frames. We used the open-source library PTLFlow [Morimitsu 2021] with the model from Morimitsu et al. [2025] for all optical flow measurements. Our method achieves a drift rate of 0.75 mm/s for reconstructed regions, significantly lower

than the 1.5 mm/s reported by Halimi et al. [2022b] as shown in Table 1, demonstrating substantially higher tracking accuracy.

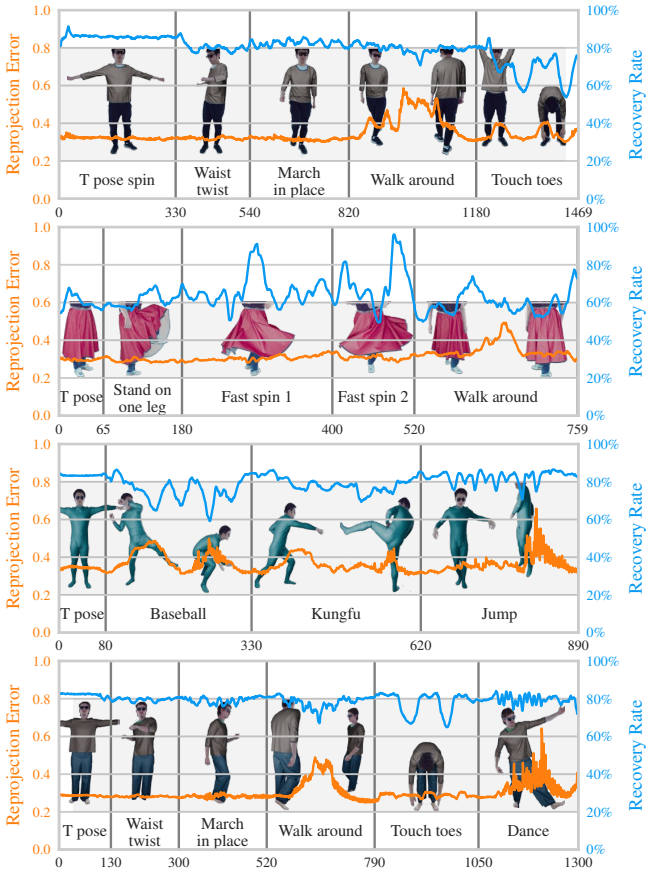


Fig. 6. *Recovery rate (blue) and 95th percentile reprojection error in image space pixels (orange) per frame for all motion sequences. Even as recovery rate varies across motions and garments, the reprojection error for recovered crosspoints remains below 0.7 pixels.*

2D Reprojection Error. 2D reprojection error is calculated by reprojecting the triangulated 3D points back into image space using the camera parameters and measuring the error against the 2D localization. This error for all four captures is presented in Fig. 6. Our 95% percentile reprojection error per frame remains below the sequence-level 95% error of 0.6979 pixels reported by Chen et al. [2021]. Note that the exact reprojection error depends on the accuracy of the camera parameters used as well as the motion sequence; for example, the spike in error during our walk-around sequences is likely due to the subject moving closer to some cameras, which increases the sensitivity to small localization errors and produces larger reprojection deviations. Nonetheless, both systems use the same image resolution for these calculations, meaning our system's consistently lower reprojection error across different frames suggests our pipeline produces more accurate localization results.

Recovery Rate. We additionally report recovery rate in Fig. 6. While again, this metric is impacted by the capture setup and motion sequence, as well as garment type, we can see that the recovery rate for the shirt and suit remains in the range of 54% to 90% reported for the shirt captured by Halimi et al. [2022b]. In contrast, the skirt shows a lower average recovery rate of approximately 60%, primarily due to the increased number of wrinkles and severe self-occlusions. The recovery rate fluctuates with the type of motion. For instance, during the marching-in-place motion, the recovery rate for the shirt varies due to the arms periodically swinging close to the torso. Similarly, during the fast-spinning motion with the skirt, the garment expands outward, revealing more visible area to the cameras and resulting in noticeable peaks in the recovery rate.

Table 2. *Image-space optical flow calculated from the composite in Fig. 9 to the input frame for triangulated and inpainted pixels.*

Garments	Triangulated	Inpainted	Overall
Skirt	0.88 px	1.25 px	0.92 px
Shirt	0.35 px	0.82 px	0.38 px
Pants & Shirt	0.39 px	0.90 px	0.43 px
Suit	0.27 px	1.16 px	0.33 px

Image-Space Optical Flow. To evaluate our full reconstruction, we perform optical flow between select camera frames and overlays of the reconstructed garment textured with the capture (see Fig. 9). Note that optical flow is not meant for comparing images depicting very different geometries. Thus, it can miss drastic differences such as the bottom hem of the shirt and the fabric fold in the left pants leg. However, it allows us to determine that triangulated vertices have subpixel drift magnitudes while the average inpainted vertex remains relatively coherent with the input; see Table 2. Qualitatively, we see that the highest optical flow magnitudes occur when large neighborhoods are inpainted. Despite this, the high error regions are relatively localized, suggesting most inpainted regions are relatively coherent with the ground truth geometry.

Using this metric, we compare against Chen et al. [2021] on the suit garment. Our method achieves 0.71 px at the 95th percentile and 1.24 px at the 99th percentile, compared to 1.20 px and 2.46 px reported by Chen et al. [2021]. While sensitivity to reconstruction rendering, optical flow network, and capture setting precludes a perfectly controlled comparison, the margin is substantial.

Synthetic Reconstruction. By rendering synthetic motion sequences using the same camera parameters as our real capture setup, we can produce synthetic multi-view images to feed into our pipeline to obtain a reconstructed mesh sequence. We can then directly compare the reconstructed 3D position of each fine marker against its ground truth position. For this experiment, we drive the suit garment with a baseball pitching motion from the CMU Graphics Lab Motion Capture Database [Carnegie Mellon University [n. d.]] using a cloth simulator [Lu et al. 2025] to produce a ground-truth mesh sequence. Compared against the ground-truth vertex positions, our full pipeline achieves 0.67 mm average per-vertex error across the sequence.

In addition, we repeat the experiment under a coarse-only setting: only the sparse L-shaped marker corners are triangulated, and the remaining dense vertices are linearly interpolated from this coarse result. This mimics the resolution of the checkerboard markers used by Chen et al. [2021], Evaluated over all vertices, this setting yields 1.10 mm average error, a 64% increase over our full pipeline. This

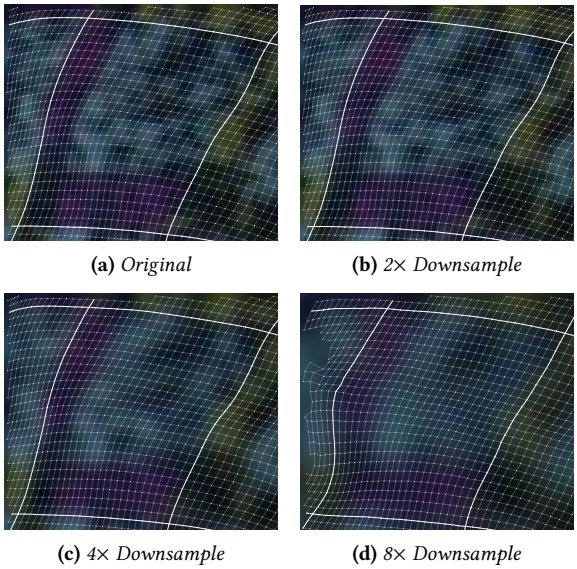
Table 3. *Pipeline robustness to image blur was evaluated by downsampling input images according to ↓ Rate. Compared to the full-resolution input, the percentage of crosspoints recovered and pixel error degrades gracefully with increasing amounts of blur.*

↓ Rate	Recovery	Mean	Median	95%	99%
2×	99%	0.06	0.04	0.18	0.41
4×	97%	0.24	0.18	0.63	1.14
8×	76%	0.77	0.65	1.82	2.71

demonstrates the significant improvement in reconstruction quality enabled by our two-level pattern and crosspoint localizer, which allows us to capture fine geometric details that are missed when relying solely on the coarse marker.

5.2 Pattern Robustness

We evaluated the robustness of our marker pattern under image degradation caused by blur, which may result from motion blur, defocus, or other factors. Specifically, we downsampled the original image and then upsampled it back to the original resolution using Lanczos interpolation with the OpenCV library. We then applied our full localization pipeline to the blurred results. A visual comparison is presented in Fig. 7, and Table 3 reports the recovery rate relative to the original resolution, along with the localization error compared to the original full-resolution image.



(a) *Original* **(b)** *2× Downsample*

(c) *4× Downsample* **(d)** *8× Downsample*

Fig. 7. *Images can be downsampled up to a factor of eight before crosspoint localization significantly degrades.*

The results show that our pattern and pipeline maintain a high recovery rate and precise localization performance up to 4× downsampling. Even under 8× downsampling, the system still achieves an acceptable recovery rate and localization error, demonstrating the robustness of our design.

Table 4. *The Marker Localizer reliably detects regions of the image that contain a marker center (top) and outputs the positions of the marker's four corners and center with sub-pixel accuracy (bottom).*

Confusion matrix	Positive		Negative	
Predicted Positive	1.964%		0.187%	
Predicted Negative	0.223%		97.63%	
Localization error	Mean	Median	95%	99%
Center	0.410	0.310	0.984	1.946
Top Left	0.492	0.355	1.238	2.692
Top Right	0.493	0.354	1.245	2.755
Bottom Left	0.498	0.360	1.260	2.722
Bottom Right	0.493	0.355	1.255	2.684

Table 5. *The crosspoint localizer accurately predicts whether a crosspoint is valid (top) and outputs its position in the canonical marker space. Its localization error in image space pixels is much lower than a similar network that directly operates on the image space (bottom).*

Confusion matrix	Positive		Negative	
Predicted Positive	96.80%		0.459%	
Predicted Negative	0.177%		2.559%	
Localization Error	Mean	Median	95%	99%
W/ Homography	0.075	0.060	0.181	0.324
W/o Homography	0.403	0.323	0.962	1.673

5.3 Neural Networks

We split the synthesized dataset into training, validation, and test sets. Our models were trained on the augmented training set for multiple epochs, and the checkpoint with the highest validation accuracy was selected for evaluation. Below, we report the performance of our three networks on the final test set.

Marker Localizer. In Table 4, we summarize the performance of the Marker Localizer. The overall classification accuracy reaches 99.594%, with a precision of 91.3% and a recall of 89.8%. In practice, we apply a threshold slightly below 0.5 during inference to favor higher recall by introducing more false positives and reducing false negatives, as the false positives will be filtered out in subsequent processing stages. While the localization accuracy is relatively low, with a mean error of 0.41 pixels for the center and 0.496 pixels for the corners, it is still sufficient for projecting the detected marker onto a canonical square. The final localization of the crosspoints, as shown later, is significantly more precise.

Marker Recognizer. Marker recognition accuracy is consistently high across all four garments, with the suit having the highest accuracy of 99.97%, and the shirt having the lowest accuracy of 99.82%. Initially, we trained separate recognizers for the suit, shirt, and skirt. For the pants, however, we reused the same set of markers as the skirt and applied the recognizer trained on skirt data. Notably, we did not preserve the marker connectivity in the pants. Despite this, we achieved high classification accuracy across all garments. This demonstrates the generalizability of our marker recognizer architecture and suggests our pipeline could be further generalized to a single trained network that can be directly applied to multiple garment types without retraining.

Table 6. *Average computation time per frame for each stage in the reconstruction pipeline.*

Garment	# Markers	# Vertices	Marker Localizer	Marker Recognizer	Crosspoint Localizer	Triangulation	Post processing
Shirt	1411	0.9M	1.76 s	1.12 s	6.72 s	14.42 s	111.48 s
Skirt	3441	2.1M	1.93 s	1.74 s	8.91 s	28.31 s	248.79 s
Suit	2190	1.5M	1.79 s	2.12 s	11.49 s	27.48 s	198.09 s
Pants & Shirt	3096	2.0M	2.21 s	3.79 s	14.07 s	36.34 s	251.63 s

Crosspoint Localizer. Our Crosspoint Localizer reaches a classification accuracy of 99.36% when identifying valid crosspoints, indicating strong reliability (see Table 5). The average localization error is 0.075 pixels in the original image space. This extremely low error demonstrates the pipeline's high precision, which is critical for the subsequent 3D reconstruction stage.

5.4 Ablation Study

To validate the effectiveness of our two-level pattern design and the corresponding localization pipeline, we conducted ablation studies on two degraded configurations: (1) using only the coarse marker, and (2) using only the dense noise pattern.

For the first experiment, we only used our pipeline's Marker Localizer and Recognizer to localize the coarse marker corners. This setup is analogous to the checkerboard pattern with symbols employed by Chen et al. [2021], where the reconstructed mesh resolution is inherently limited by the sparse marker density. Indeed, we can examine the reconstruction results in Fig. 8 and see that the coarse-only approach fails to capture fine geometric details. A denser pattern is necessary for high-resolution reconstruction. As mentioned in Sec. 5.1, the coarse-only approach yields 1.10 mm average per-vertex error on synthetic data, compared to 0.67 mm for our full pipeline.

(a) *Two-Level Pattern* **(b)** *Coarse Only Pattern*

Fig. 8. *Using only the sparse markers for reconstruction results in much lower resolution than using the full pipeline with crosspoints. Figures use flat shading to emphasize the difference.*

In this second experiment, we directly used the dense noise pattern for localization, forgoing the homography transformation provided by the coarse marker. We trained a modified crosspoint localizer that takes resized crops without homography rectification as well as a reference texture as input. As shown in the last row of Table 5, the dense-only approach exhibits significantly larger localization error than the full pipeline. This shows the importance of the homography rectification enabled by the coarse marker, which allows the dense localizer to focus on precise subpixel localization within a canonical coordinate frame. The synergy between these two levels is key to achieving both high precision and high resolution in our capture system.

5.5 Setup and Computation Time

The computation time for each component of our pipeline is reported in Table 6. Note that our implementation is not fully optimized, so these timings should be treated as reference values. We use an NVIDIA RTX Ada 6000 GPU for neural network inference and physics-based post-processing, and an AMD Ryzen Threadripper PRO 5975WX CPU for triangulation.

The primary bottleneck is the physics-based post-processing: VBD is designed for scenarios with many small objects each containing thousands of vertices, whereas our garments are single connected meshes with up to 2.1M vertices (the skirt). In this regime, global error reduction requires many iterations, as local solvers struggle to propagate corrections across the full domain. The multi-resolution strategy mitigates this, but still requires running the optimization multiple times per frame at each resolution level. Like Chen et al. [2021] and Halimi et al. [2022a], our pipeline targets offline data collection rather than real-time applications, and reconstruction quality is prioritized over runtime. Exploring how the post-processing step could be made more efficient despite the high-resolution meshs required for high-quality reconstruction would be an interesting future direction to explore.

6 Discussion and Future Directions

We have proposed a two-level marker pattern and corresponding localization pipeline for high-fidelity cloth capture, achieving a spatial resolution of 1 mm that preserves fine details such as wrinkles and folds. Combined with a physics-based refinement pipeline, our system enables high-quality 4D cloth capture over time. This work allows for high-resolution capture of garment deformation even in capture setups with few cameras. However, many interesting future directions exist for further improving the usability of our pipeline.

Our captured sequences generally have a low recovery rate at the seams of the garments. This is partially due to fabrication inconsistencies in how the seams are lined up as well as non-patterned elements such as zippers interrupting the cloth. While a skilled sewist can reduce these artifacts, the primary reason for low recovery in these regions is that our pipeline uses square markers of a fixed size. This means that full markers cannot fit at the seams. Furthermore, panels that are thin or irregularly shaped, such as

Fig. 9. *Challenging examples from all four capture sequences. In addition to the input image and full reconstruction (left two columns), we show the optical flow between the input image and retextured reconstruction (third column) and the highlighted inpainted regions (fourth column) for comparison. Color hue denotes optical flow angle, while intensity denotes magnitude.*

the waistband of the skirt or the collar of the suit, cannot be reconstructed. Extending our pipeline to support other marker shapes is one way to resolve this issue.

Another way to address this problem would be to improve our pipeline's ability to localize crosspoints in partial markers. Currently, our pipeline identifies and transforms incomplete markers using information from adjacent full markers. This strategy worsens the quality of the homography transformation, and it cannot identify partial markers whose neighbors are occluded. Supporting partial marker detection would improve robustness and accuracy at garment boundaries as well as regions with significant self-occlusion.

Our reconstruction method could also be extended to make better use of all captured data. Right now, triangulation is only performed for vertices captured by at least two cameras. This is because a single camera view alone is insufficient to provide unambiguous 3D localization. However, a single view still constrains the set of plausible cloth positions. More closely coupling our triangulation and 3D post-processing as a single optimization problem could allow for this partial information to be included. Triangulated regions could also be used to infer cloth parameters in order to improve the physics-guided inpainting. Estimating cloth parameters from 4D capture would not only improve the post-processing part of our pipeline, but it would also benefit cloth modeling, simulation, and material parameter estimation [Feng et al. 2022; Larionov et al. 2022; Rodriguez-Pardo et al. 2023].

Finally, while our pipeline is focused on reconstructing cloth in motion, we believe our insights about two-level pattern localization and physics-based post-processing could be extended to the reconstruction of other elastic bodies. Humans are constantly interacting with soft, deformable objects that we struggle to accurately model. Our hope is that this work can be used to create high-quality, diverse, and large-scale datasets of deformable objects by enabling high-fidelity 4D reconstruction in resource-constrained applications.

Acknowledgments

The synthetic motion data used in this project was obtained from mocap.cs.cmu.edu, which is a database created with funding from NSF EIA-0196217. The authors would also like to thank Colin Galbraith for being the motion capture subect for Pants & Shirt. Yin Yang was supported in part by NSF grant #2301040.

References

Eunsoo Baek, Shelley Haines, Omar H Fares, Zhihong Huang, Yuwei Hong, and Seung Hwan Mark Lee. 2022. Defining digital fashion: Reshaping the field via a systematic review. *Computers in Human Behavior* 137 (2022), 107407.

Nicoló Biasi, Francesco Setti, Alessio Bue, Mattia Tavernini, Massimo Lunardelli, Alberto Fornaser, Mauro Lio, and Mariolino Cecco. 2015. Garment-Based Motion Capture (GaMoCap): High-Density Capture of Human Shape in Motion. *Mach. Vision Appl.* 26, 7-8 (Nov. 2015), 955–973. doi:10.1007/s00138-015-0701-2

Derek Bradley, Tiberiu Popa, Alla Sheffer, Wolfgang Heidrich, and Tamy Boubekeur. 2008. Markerless Garment Capture. In *ACM SIGGRAPH 2008 Papers (SIGGRAPH '08)*. Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/1399504.1360698

Carnegie Mellon University. [n. d.]. CMU MoCap Dataset.

Anka He Chen, Jerry Hsu, Ziheng Liu, Miles Macklin, Yin Yang, and Cem Yuksel. 2025. Offset Geometric Contact. *ACM Trans. Graph.* 44, 4, Article 160 (July 2025), 21 pages. doi:10.1145/3731205

Anka He Chen, Ziheng Liu, Yin Yang, and Cem Yuksel. 2024. Vertex Block Descent. *ACM Trans. Graph.* 43, 4 (July 2024), 116:1–116:16. doi:10.1145/3658179

He Chen, Hyojoon Park, Kutay Macit, and Ladislav Kavan. 2021. Capturing Detailed Deformations of Moving Human Bodies. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–18.

Xiaowu Chen, Bin Zhou, Feixiang Lu, Lin Wang, Lang Bi, and Ping Tan. 2015. Garment Modeling with a Depth Camera. *ACM Trans. Graph.* 34, 6 (Nov. 2015), 203:1–203:12. doi:10.1145/2816795.2818059

Toby Chong, I-Chao Shen, Nobuyuki Umetani, and Takeo Igarashi. 2021. Per Garment Capture and Synthesis for Real-time Virtual Try-on. In *The 34th Annual ACM Symposium on User Interface Software and Technology (UIST '21)*. Association for Computing Machinery, New York, NY, USA, 457–469. doi:10.1145/3472749.3474762

R. Daněřek, E. Dibra, C. Öztireli, R. Ziegler, and M. Gross. 2017. DeepGarment : 3D Garment Shape Estimation from a Single Image. *Computer Graphics Forum* 36, 2 (2017), 269–280. doi:10.1111/cgf.13125

Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. 2008. Performance Capture from Sparse Multi-View Video. In *ACM SIGGRAPH 2008 Papers (SIGGRAPH '08)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/1399504.1360697

Wei-Wen Feng, Yizhou Yu, and Byung-Uck Kim. 2010. A Deformation Transformer for Real-Time Cloth Animation. *ACM Trans. Graph.* 29, 4 (July 2010), 108:1–108:9. doi:10.1145/1778765.1778845

Xudong Feng, Wenchao Huang, Weiwei Xu, and Huamin Wang. 2022. Learning-Based Bending Stiffness Parameter Estimation by a Drape Tester. *ACM Trans. Graph.* 41, 6 (Nov. 2022), 221:1–221:16. doi:10.1145/3550454.3555464

Eitan Grinspun, Anil N Hirani, Mathieu Desbrun, and Peter Schröder. 2003. Discrete shells. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*. 62–67.

Jingfan Guo, Fabian Prada, Donglai Xiang, Javier Romero, Chenglei Wu, Hyun Soo Park, Takaaki Shiratori, and Shunsuke Saito. 2023. Diffusion Shape Prior for Wrinkle-Accurate Cloth Registration. arXiv:2311.05828 [cs.CV] https://arxiv.org/abs/2311.05828

Oshri Halimi, Fabian Prada, Tuur Stuyck, Donglai Xiang, Timur Bagautdinov, He Wen, Ron Kimmel, Takaaki Shiratori, Chenglei Wu, and Yaser Sheikh. 2022a. Garment Avatars: Realistic Cloth Driving Using Pattern Registration. doi:10.48550/arXiv.2206.03373 arXiv:2206.03373 [cs]

Oshri Halimi, Tuur Stuyck, Donglai Xiang, Timur Bagautdinov, He Wen, Ron Kimmel, Takaaki Shiratori, Chenglei Wu, Yaser Sheikh, and Fabian Prada. 2022b. Pattern-Based Cloth Registration and Sparse-View Animation. *ACM Trans. Graph.* 41, 6 (Nov. 2022), 196:1–196:17. doi:10.1145/3550454.3555448

Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. 2024. GaussianAvatar: Towards Realistic Human Avatar Modeling from a Single Video via Animatable 3D Gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 634–644.

G. Irving, J. Teran, and R. Fedkiw. 2004. Invertible finite elements for robust simulation of large deformation. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Grenoble, France) *(SCA '04)*. Eurographics Association, Goslar, DEU, 131–140. doi:10.1145/1028523.1028541

Tasin Islam, Alina Miron, Xiaohui Liu, and Yongmin Li. 2024. Deep Learning in Virtual Try-On: A Comprehensive Survey. *IEEE Access* 12 (2024), 29475–29502. doi:10.1109/ACCESS.2024.3368612

Egor Larionov, Marie-Lena Eckert, Katja Wolff, and Tuur Stuyck. 2022. Estimating Cloth Elasticity Parameters Using Position-Based Simulation of Compliant Constrained Dynamics. doi:10.48550/arXiv.2212.08790 arXiv:2212.08790 [cs]

Bo Li, Lionel Heng, Kevin Koser, and Marc Pollefeys. 2013. A Multiple-Camera System Calibration Toolbox Using a Feature Descriptor-Based Calibration Pattern. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 1301–1307. doi:10.1109/IROS.2013.6696517

Rong-Hao Liang, Hannah van Iterson, Holly Krueger, Marina Toeters, and Loe Feijs. 2024. Chic-Marker: Fashionably Fusing Fiducial Markers into Apparel and Accessories. In *Proceedings of the 9th ACM Symposium on Computational Fabrication (SCF '24)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3639473.3665790

Lijuan Liu, Xiangyu Xu, Zhijie Lin, Jiabin Liang, and Shuicheng Yan. 2023. Towards Garment Sewing Pattern Reconstruction from a Single Image. *ACM Trans. Graph.* 42, 6 (Dec. 2023), 200:1–200:15. doi:10.1145/3618319

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graph.* 34, 6 (Oct. 2015), 248:1–248:16. doi:10.1145/2816795.2818013

Zixuan Lu, Ziheng Liu, Lei Lan, Huamin Wang, Yuko Ishiwaka, Chenfanfu Jiang, Kui Wu, and Yin Yang. 2025. High-performance CPU Cloth Simulation Using Domain-decomposed Projective Dynamics. *ACM Trans. Graph.* 44, 4, Article 51 (July 2025), 17 pages. doi:10.1145/3731182

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5442–5451.

Gyeongsik Moon, Hyeongjin Nam, Takaaki Shiratori, and Kyoung Mu Lee. 2022. 3D Clothed Human Reconstruction in the Wild. In *Computer Vision – ECCV 2022*,

Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 184–200. doi:10.1007/978-3-031-20086-1_11

Henrique Morimitsu. 2021. PyTorch Lightning Optical Flow. https://github.com/hmorimitsu/ptlflow.

Henrique Morimitsu, Xiaobin Zhu, Roberto M. Cesar, Xiangyang Ji, and Xu-Cheng Yin. 2025. DPFlow: Adaptive Optical Flow Estimation with a Dual-Pyramid Framework. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 17810–17820. doi:10.1109/CVPR52734.2025.01659

Newton Contributors. 2025. *Newton: GPU-accelerated physics simulation for robotics, and simulation research*. Newton a Series of LF Projects, LLC. https://github.com/newton-physics/newton

Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. 2020. Tailornet: Predicting Clothing in 3d as a Function of Human Pose, Shape and Garment Style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7365–7375.

Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J. Black. 2017. ClothCap: Seamless 4D Clothing Capture and Retargeting. *ACM Trans. Graph.* 36, 4 (July 2017), 73:1–73:15. doi:10.1145/3072959.3073711

T. Popa, Q. Zhou, D. Bradley, V. Kraevoy, H. Fu, A. Sheffer, and W. Heidrich. 2009. Wrinkling Captured Garments Using Space-Time Data-Driven Deformation. *Computer Graphics Forum* 28, 2 (2009), 427–435. doi:10.1111/j.1467-8659.2009.01382.x

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 779–788.

Carlos Rodriguez-Pardo, Melania Prieto-Martin, Dan Casas, and Elena Garces. 2023. How Will It Drape Like? Capturing Fabric Mechanics from Depth Images. *Computer Graphics Forum* 42, 2 (2023), 149–160. doi:10.1111/cgf.14750

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. 2019. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2304–2314.

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 84–93.

Igor Santesteban, Miguel A. Otaduy, and Dan Casas. 2019. Learning-Based Animation of Clothing for Virtual Try-On. *Computer Graphics Forum* 38, 2 (2019), 355–366. doi:10.1111/cgf.13643

Volker Scholz, Timo Stich, Marcus Magnor, Michael Keckeisen, and Markus Wacker. 2005. Garment Motion Capture Using Color-Coded Patterns. In *ACM SIGGRAPH 2005 Sketches (SIGGRAPH '05)*. Association for Computing Machinery, New York, NY, USA, 38–es. doi:10.1145/1187112.1187157

Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. 2008. Articulated Mesh Animation from Multi-View Silhouettes. In *ACM SIGGRAPH 2008 Papers (SIGGRAPH '08)*. Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/1399504.1360696

Huamin Wang, Florian Hecht, Ravi Ramamoorthi, and James F. O'Brien. 2010. Example-Based Wrinkle Synthesis for Clothing Animation. In *ACM SIGGRAPH 2010 Papers (SIGGRAPH '10)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/1833349.1778844

Ryan White, Keenan Crane, and D. A. Forsyth. 2007. Capturing and Animating Occluded Cloth. *ACM Trans. Graph.* 26, 3 (July 2007), 34–es. doi:10.1145/1276377.1276420

Shan Yang, Zherong Pan, Tanya Amert, Ke Wang, Licheng Yu, Tamara Berg, and Ming C. Lin. 2018. Physics-Inspired Garment Recovery from a Single-View Image. *ACM Trans. Graph.* 37, 5 (Nov. 2018), 170:1–170:14. doi:10.1145/3026479