# XR360: A TOOLKIT FOR MIXED 360 AND 3D PRODUCTIONS

*Antonis Karakottas, Nikolaos Zioulis, Alexandros Doumanglou, Vladimiros Sterzentsenko*
*Vasileios Gkitsas, Dimitrios Zarpalas and Petros Daras*

Centre for Research and Technology Hellas (CERTH) - Information Technologies Institute (ITI),
Thessaloniki, Greece
{ankarako, nzioulis, aldoum, vladster, gkitsav, zarpalas, daras}@iti.gr

## ABSTRACT

Due to the affordability and high quality of the emerging XR devices and platforms, immersive media content has become more accessible to a wide variety of consumers while their technology gains increased interest from content creators. However, the previously mentioned technological advancements are focused more on sensor or rendering-display technologies, while the development of software tools that assist professionals, indies or hobbyists to create immersive content and narratives is left behind. In this work we introduce XR360, a unified toolkit for producing $360^o$ immersive experiences by embedding 3D captured human performances in $360^o$ captured scenes. The presented toolkit comprises a complete solution for content acquisition, 3D asset creation and 3D-$360^o$ fusion. It combines a volumetric capturing studio, a performance capture production tool and a $360^o$ fusion application, designed to operate in a cascaded, but disjoint, manner while remaining user-friendly by hiding complex technicalities from its end-users.

***Index Terms***— Immersive Media, Volumetric Video, Performance Capture, Omnidirectional Media, Mixed Production, VR, 3D, RGB-D

## 1. INTRODUCTION

The rapid development of immersive media technologies has significantly changed the way people can experience multimedia content. Novel and diverse platforms have emerged that can engage consumers and place them in virtual or augmented environments, giving them the ability to even interact with each other, and therefore increasing their sense of presence while being affordable and easy to use. At the same time, producers and content creators have access to a variety of tools that help them create and distribute immersive stories that convey narratives and experiences, having a high impact on different industry sectors such as movies [1] and video games, or educational training and medicine [2], architecture [3], data visualization [4], rehabilitation [5] or cultural heritage [6].

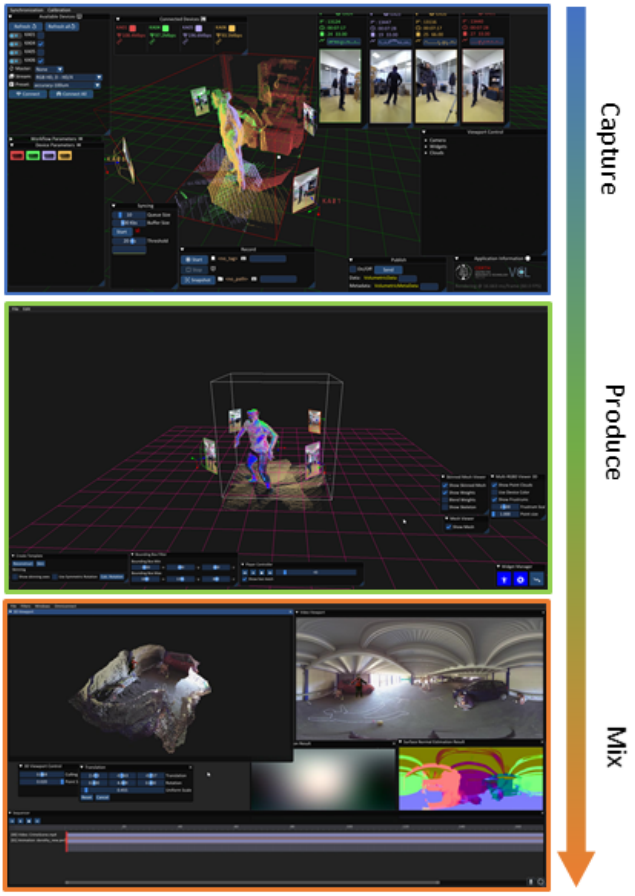


**Fig. 1**. Screenshots of the three cascaded XR360 toolkit. **Top:** The volumetric capture tool with a live point cloud visualization of the actor in the capturing space. **Middle:** The performance capture application while tracking the movement of the actor. **Bottom:** The $360^o$ Fusion tool, visualizing a $360^o$ scene's estimated point cloud, surface normal and illumination.

Capturing and producing immersive content often-most involves blending the "real" with the "virtual" and requires

the integration of a wide variety of different sensor technologies as well as media formats such as 3D animated graphics, audio, $360^o$ or multi-view video. While a variety of tools for content acquisition and production are available, the process of developing immersive multimedia content remains challenging as it usually requires high technological knowledge regarding the hardware and the software that will be used.

This challenge is even more pronounced for volumetric video production pipelines. Volumetric video adds hardware complexity (due to its requirement for numerous sensors) as well as a lack of mature software (given its relatively recent emergence). Typically, volumetric capturing studios that produce free viewpoint, 3D content, require a sizeable capturing space, a technically non-trivial extrinsic sensor calibration procedure and many different sensor devices (cameras, infrared and/or depth/stereo pairs) to be integrated and synchronized [7, 8]. Another important issue is that most of the existing solutions do not produce ready-to-use 3D assets that integrate well with existing production and editing tools.

Most immersive (or mixed) productions now rely on standard and established production tools. These, however, are complex applications that need to support a variety of use cases. More specialized variants like Wonda VR[1] and Cinema4D[2] focus on improved and focused support for specific immersive media types but fall short on supporting recent technological advances. Further, their 3D compositing capacity is limited to synthetic objects, limiting their narrations to either directed captures or computer generated avatars.

In this paper we present a complete toolkit that supports end-to-end mixed immersive content production workflows. By mixed we refer to disjoint scene and performer captures that is enabled by the integration of different technologies. Volumetric video capturing of human performances digitizes real performers and allows for their compositing into in $360^o$ media. Our toolkit covers all aspects of such productions by providing an end-to-end pipeline for 3D capturing, animated 3D asset creation and 3D-$360^o$ fusion, leaving only $360^o$ captures as a separate action, which however, is now relatively mature. The XR360 toolkit comprises:

1. a portable and affordable hardware and software **volumetric capturing tool** which utilizes the latest consumer level RGB-D sensor technology,

2. a **3D production tool** based on an advanced human digitization technology that converts recorded multi-view RGB-D sequences of human performances to animated 3D assets, and, finally,

3. a **mixed media (3D and $360^o$) fusion tool** that allows for the realistic compositing of animated 3D objects in $360^o$ videos.

Sample screenshots of these tools are offered in Fig 1.

## 2. RELATED WORK

Many software and hardware tools for immersive content creation have been made available in the last decade. These span from applications that target realistic Computer Generated (CG) object rendering in real scenes [9, 10] and immersive data visualization [4], to interactive and personalized narrative creation for immersive virtual environments [11] or complete frameworks with many utilities for creating immersive experiences. These works focus on specific aspects of immersive content like the compositing of objects within $360^o$ content or how to better facilitate its storytelling.

Another body of work focuses on creating frameworks that help creators and developers in building XR applications. The VREX [12] framework allows for creating gaming VR applications on the HTC Vive[3] platform and is developed in the Unity game engine[4]. VREX provides a variety of white-label solutions for player-scene interactions, therefore enabling fast prototyping and developing of VR applications. Similarly, VR-Rides [13] is a framework with ready-made components that assist in VR health application development, utilizing physical movement sensors and Google Street View $360^o$ images. Furthermore, the authors of [14] present an authoring tool for creating educational applications, targeting mobile platforms. The presented framework utilizes $360^o$ media with 3D elements and 3D audio and is showcased through a marine biology educational fishing game.

Recently many solutions emerged that tackle realistic CG object rendering in real scenes captured using traditional RGB or $360^o$ cameras, in order to increase the immersion of the user by being placed in augmented physical environments. Specifically, in [9], an application developed also in Unity is presented, in which CG models are efficiently embedded and rendered in conventional Low Dynamic Range (LDR) $360^o$ videos. This is achieved by first converting the LDR video via inverse tone mapping and detecting the most salient lights in the scene. The framework utilizes image based lighting and shadowing to efficiently shade the CG objects added in the $360^o$ video. Moreover, in [10] the authors present a system for rendering CG objects in underwater $360^o$ scenes utilizing similar techniques but further enhancing them with complex lighting effects that combine underwater caustics, god rays, fog and particles developed on the Unreal gaming engine[5].

Finally, full-grown authoring frameworks that bring immersive content creation a step closer to the non-professionals have been released, with WondaVR and Forge.js[6] being two representative examples of such applications. These frameworks offer a wide variety of tools for creating immersive ex-

[1] https://www.wondavr.com/

[2] https://www.maxon.net/en-us/products/cinema-4d/overview/

[3] https://www.vive.com/us/

[4] https://unity.com/

[5] https://www.unrealengine.com/en-US/

[6] https://forgejs.org/

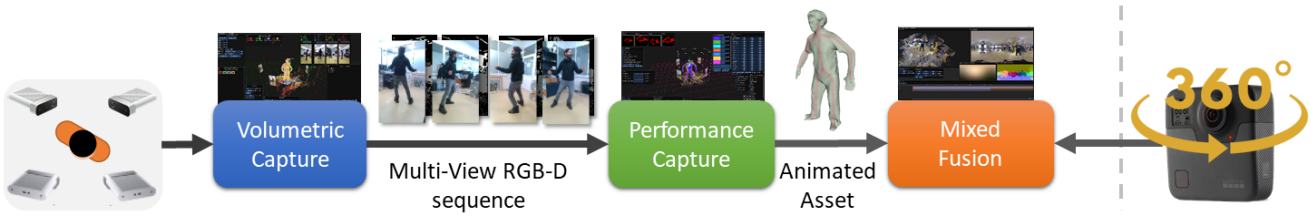


**Fig. 2**. A high-level overview of the 360D Studio's workflow. The actor's performance is captured by the volumetric capture application which produces a multi-view RGB-D sequence. The actor's performance is tracked by the perfomance capture tool and exported into an animated 3D mesh, which can be fused in an already existing $360^o$ production.

periences from $360^o$ media by enriching them with interactive hot-spots and 3D elements.

Evidently, there are no tools for merging the two different variants of free viewpoint video formats, namely volumetric and omnidirectional. While some tools offer 3D composition capabilities, they do not allow for realistic captures, and usually do not offer high quality rendering for the composited content. Last, AI integration is limited, while it currently is the most promising technology for alleviating production workloads.

## 3. XR360

The XR360 toolkit is designed for augmenting $360^o$ media with 3D elements. It provides the ability to capture real human actor performances in 3D and fuse these performances with a $360^o$ capture, resulting a new $360^o$ video where the viewer's experience maybe improved as he or she can interact with the newly inserted digitized humans. Fig. 2 presents a high-level overview of the toolkit's workflow, starting with multi-view capturing and followed by 3D animated asset creation, which can be embedded in already existing $360^o$ productions.

In this section, we present the software and hardware components of the toolkit. At first, we describe the multi-view acquisition system which exploits the latest advancements in commercial-grade RGB-D sensor technology. Subsequently, we present the animated 3D human asset creation tool that integrates a fully automated performance capture technology and finally, we showcase the 360 fusion tool which employs modern AI capabilities in $360^o$ scene understanding in order to guide the process of compositing the human 3D assets within $360^o$ content.

### 3.1. Capture

The acquisition system [15] is responsible for the volumetric capture of the actor's performance[7]. It is a multi-view system developed in a distributed client-server architecture, where server software is deployed on acquisition nodes with each node being composed of an RGB-D sensor and a mini-PC, while the client is deployed on a main workstation and is responsible for orchestrating the process of capturing and recording. This design allows horizontal scaling and portability while being affordable.

We utilize the latest advances in RGB-D sensor technology employing both the Intel Realsense D400 sensors [16] (specifically the D415 sensor[8]), as well as the very recently released Microsoft Azure Kinect[9]. The system is designed in a modular fashion and can thus, support an arbitrary number of sensors. Further, and most importantly, the system also is supported by the state-of-the-art method of [17][10] for volumetric sensor alignment (i.e. extrinsic sensor calibration), that leverages low-cost packaging boxes and allows arbitrary camera placement configurations.

Apart from orchestrating the capturing and recording process, in cooperation with the servers, the client component also allows for individual parameterization of each sensor in order to control the capturing conditions (like camera exposure time, depth sensor's laser power, stream resolution, stream compression settings, etc. ) and achieve higher quality recordings.

### 3.2. Produce

Following the recording of an actor performance, the 3D production tool transforms the volumetrically aligned RGB-D streams into an animated 3D asset, which can be used for 3D-$360^o$ fusion, or any other purpose. It employs automated performance capture technology, which requires no external photogrammetry or rigging software as is typical for other systems. Apart from lowering the cost and technical barrier, this also smooths the workflow. The template creation part that involves creating the actor's 3D model and associated metadata required for animating it (bone hierarchy and skinning weights), albeit optional, is integrated and automated into the

[7] https://github.com/VCL3D/VolumetricCapture

[8] https://www.intelrealsense.com/depth-camera-d415/

[9] https://azure.microsoft.com/en-gb/services/kinect-dk/

[10] https://vcl3d.github.io/StructureNet/

tool. It is also possible to reuse templates among captures, or even use an externally provided one.

We extend the existing method of [18] by automating template model creation, optimizing its run-time through warp aggregated parallel operations, while also focusing on the more robust optimization layer. In more detail, the automated template creation pipeline involves a Poisson reconstruction [19] of a volumetric snapshot, followed by Pinnochio skinning [20]. This actor template model creation process is presented in Fig. 3.

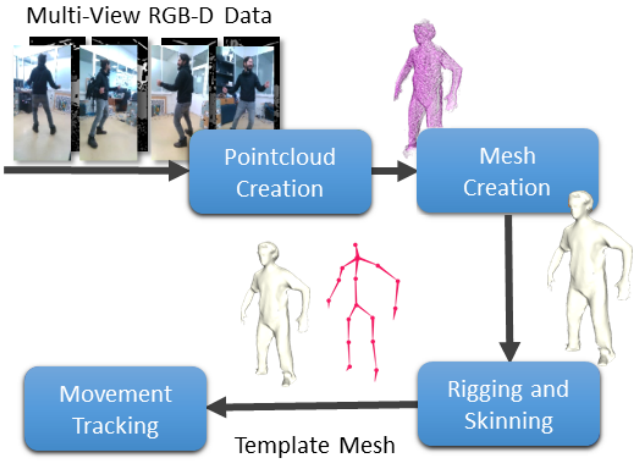


**Fig. 3**. Performance Capture initialization. In order to start tracking the actor's movements in the recorded multi-view sequences, first a template mesh from a reference frame must be extracted. The depth enhanced multi-view frame is converted to a registered point-cloud, from which a 3D mesh is generated using [19]. Finally, the resulting mesh is rigged and skinned by employing [20].

The multi-view output of our capturing tool is fused into a volumetric representation which is subsequently used to drive the optimization process. The latter fits the animated template model into the live data as depicted in Fig 4.

We additionally complement the standard optimization approach with two important features: manual adjustments and finer-grained optimization. The former allows for manual corrections to be applied in erroneous frames and seamlessly continue tracking without hurting the subsequent optimization steps. The latter is a modular re-design of the optimization engine to enable subsequent fine-tuning by selecting optimizer groups and associating them with different error terms and joints. The resulting output is an animated 3D model depicting the user's performance.

### 3.3. Mix

Finally, we composite the digitized performance within a $360^o$ video. Embedding 3D objects in $360^o$ videos is a challenging task that requires a lot of modalities about the scene's

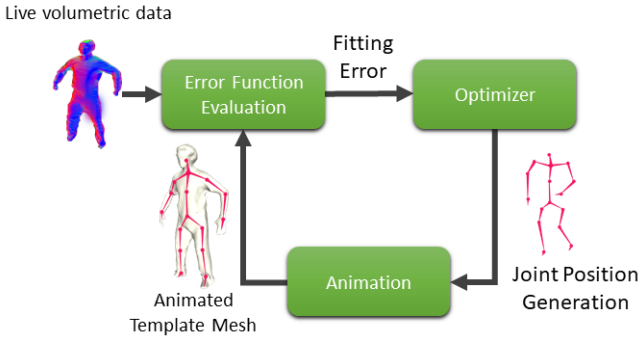


**Fig. 4**. Performance Capture procedure. First, the template mesh is initialized to a random pose by the optimizer which is animated accordingly and a distance error is calculated between the animated mesh and the live one. The error is forwarded to the optimizer which generates a pose w.r.t the incoming error until the error is minimized.

structural information to be known. To that end, our tool relies on two technologies: raytracing and AI-based scene understanding. The former is necessary for reproducing the distortions of traditional $360^o$ storage formats as traditional rasterization pipelines would achieve this through inefficient and error inducing techniques. The latter can greatly assist users with tedious trial-and-error tasks such as positioning and lighting. Understanding the scene's depth can assist with 3D object placement, while extracting scene's surface normals and illumination can aid the process of object shading.

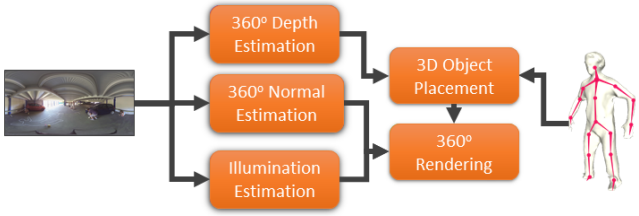


**Fig. 5**. 360 Fusion application pipeline. The $360^o$ video's depth, surface normal and illumination are estimated, to assist with 3D object placement and realistic rendering.

The fusion tool leverages recent works on data-driven depth [21, 22] and surface normal [23] estimation, as well as lighting environment map estimation.

These modalities, with an example presented in Fig. 5, can seamlessly be exploited into its rendering engine to increase the realism of the result. In addition, via a 3D user interface and a 3D visualization, content creators can naturally position the digitised content, while also receiving a real-time preview on the rendered equirectangular view.

Our rendering engine is implemented with NVidia's Op-

**Fig. 6**. Mixed production result. Top row: Sample frames from a $360^o$ video. Middle row: Sample frames of the animated asset exported by the Performance Capture tool. Bottom row: The same sample frames fused together.

tiX[11] framework and facilitates the unification of two diverse media types, 3D capture which is an inwards looking free viewpoint format, and $360^o$ which is an outwards looking free viewpoint. While the result still remains $360^o$, the freedom that 3D allows can offer content creators new opportunities for creative directing and allow for content re-use.

## 4. RESULTS AND DISCUSSION

An exemplary mixed production is presented in Fig. 6 with the original $360^o$ content depicting a gym commercial (top). An actor was 3D captured performing jumping jacks (middle) and fused (bottom) into the original content. These type of mixed productions can decouple the storytelling and directing of spherical content from the actual shooting. In this way, content producers can overcome the challenges associated with omnidirectional productions and gain an added level of flexibility and enhanced creativity opportunities.

In conclusion, we presented XR360, a tookit comprising hardware and software tools for capturing multi-view videos of human performances, converting them to animated 3D assets and finally embedding them in existing $360^o$ media. Our goal is to faciliate realistic positioning and rendering. Our toolkit and workflow is designed with usability and affordability in mind and thus can be utilized by non-experts as well as professionals equally. Albeit reliance on lower-cost systems and automated processes may have an impact on the resulting quality, it is open for improvement through follow up research and development. In the near future we plan to explore data-driven approaches on the production part and also employ human guided AI during the fusion of the two media to improve performance, robustness and the resulting quality. In addition, template-less performance capture methods that utilize a parameterizable human mesh proxy such as [24] to account for quality loss of our template-based method and enable the fusion of information over time. Overall, our toolkit showcases the future for AI-enhanced mixed content productions, and the possibility of enhancing traditional captures with digitized performances. We expect that this approach can disrupt traditional VR film-making by decoupling scenic captures and storytelling.

[11] https://developer.nvidia.com/optix

**Acknowledgements:** This work was supported and received funding from the European Union Horizon H2020 Framework Programme project Hyper360, under Grant Agreement no. 761934. We are also grateful and acknowledge the support of Rundfunk Berlin-Branderrburg (RBB) and Mediaset-RTI for offering their $360^o$ productions.

## 5. REFERENCES

[1] Quentin Galvane, I-Sheng Lin, Marc Christie, and Tsai-Yen Li, "One man movie: A VR authoring tool for film previsualisation," 2016.

[2] Vuthea Chheang, Patrick Saalfeld, Bernhard Preim, Christian Hansen, Tobias Huber, Florentine Huettl, and Werner Kneist, "An interactive demonstration of collaborative vr for laparoscopic liver surgery training," in *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 2019, pp. 247–2471.

[3] Katrin Wolf, Markus Funk, Rami Khalil, and Pascal Knierim, "Using virtual reality for prototyping interactive architecture," in *Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia*, 2017, pp. 457–464.

[4] Ronell Sicat, Jiabao Li, JunYoung Choi, Maxime Cordeil, Won-Ki Jeong, Benjamin Bach, and Hanspeter

Pfister, "Dxr: A toolkit for building immersive data visualizations," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 715–725, 2018.

[5] Florian Kern, Carla Winter, Dominik Gall, Ivo Käthner, Paul Pauli, and Marc Erich Latoschik, "Immersive virtual reality and gamification within procedurally generated environments to increase motivation during gait rehabilitation," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2019, pp. 500–509.

[6] Thomas P Kersten, Felix Tschirschwitz, Simon Deggim, and Maren Lindstaedt, "Virtual reality for cultural heritage monuments–from 3d data recording to immersive visualisation," in *Euro-Mediterranean Conference*. Springer, 2018, pp. 74–83.

[7] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al., "Holoportation: Virtual 3d teleportation in real-time," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, 2016, pp. 741–754.

[8] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan, "High-quality streamable free-viewpoint video," *ACM Transactions on Graphics (ToG)*, vol. 34, no. 4, pp. 1–13, 2015.

[9] Taehyun Rhee, Lohit Petikam, Benjamin Allen, and Andrew Chalmers, "Mr360: Mixed reality rendering for 360 panoramic videos," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 4, pp. 1379–1388, 2017.

[10] Stephen Thompson, Andrew Chalmers, and Taehyun Rhee, "Real-time mixed reality rendering for underwater 360 videos," in *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2019, pp. 74–82.

[11] Rozenn Bouville, Valérie Gouranton, Thomas Boggini, Florian Nouviale, and Bruno Arnaldi, "# five: High-level components for developing collaborative and interactive virtual environments," in *2015 IEEE 8th Workshop on Software Engineering and Architectures for Realtime Interactive Systems (SEARIS)*. IEEE, 2015, pp. 33–40.

[12] Ryan Blonna, Mel Stychen Tan, Vanessa Tan, Anna Patricia Mora, and Rowel Atienza, "Vrex: A framework for immersive virtual reality experiences," in *2018 IEEE Region Ten Symposium (Tensymp)*. IEEE, 2018, pp. 118–123.

[13] Yifan Wang, Kiran Ijaz, and Rafael A Calvo, "A software application framework for developing immersive virtual reality experiences in health domain," in *2017 IEEE Life Sciences Conference (LSC)*. IEEE, 2017, pp. 37–30.

[14] Kanghyun Choi, Yeo-Jin Yoon, Oh-Young Song, and Soo-Mi Choi, "Interactive and immersive learning using 360 virtual reality contents on mobile platforms," *Mobile Information Systems*, vol. 2018, 2018.

[15] Vladimiros Sterzentsenko, Antonis Karakottas, Alexandros Papachristou, Nikolaos Zioulis, Alexandros Doumanoglou, Dimitrios Zarpalas, and Petros Daras, "A low-cost, flexible and portable volumetric capturing system," in *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2018, pp. 200–207.

[16] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik, "Intel realsense stereoscopic depth cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1–10.

[17] Vladimiros Sterzentsenko, Alexandros Doumanoglou, Spyridon Thermos, Nikolaos Zioulis, , Dimitrios Zarpalas, and Petros Daras, "Deep soft procrustes for markerless volumetric sensor alignment," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2020.

[18] Dimitrios S Alexiadis, Nikolaos Zioulis, Dimitrios Zarpalas, and Petros Daras, "Fast deformable model-based human performance capture and fvv using consumer-grade rgb-d sensors," *Pattern Recognition*, vol. 79, pp. 260–278, 2018.

[19] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe, "Poisson surface reconstruction," in *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006, vol. 7.

[20] Ilya Baran and Jovan Popović, "Automatic rigging and animation of 3d characters," *ACM Transactions on graphics (TOG)*, vol. 26, no. 3, pp. 72–es, 2007.

[21] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras, "Omnidepth: Dense depth estimation for indoors spherical panoramas," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 448–465.

[22] Nikolaos Zioulis, Antonis Karakottas, Dimitris Zarpalas, Federic Alvarez, and Petros Daras, "Spherical view synthesis for self-supervised $360^o$ depth estimation," in *International Conference on 3D Vision (3DV)*, September 2019.

[23] Antonis Karakottas, Nikolaos Zioulis, Stamatis Samaras, Dimitrios Ataloglou, Vasileios Gkitsas, Dimitrios Zarpalas, and Petros Daras, "360 surface regression with a hyper-sphere loss," in *International Conference on 3D Vision*, September 2019.

[24] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *European Conference on Computer Vision*. Springer, 2016, pp. 561–578.