# 4D Visualization of Dynamic Events from Unconstrained Multi-View Videos

Aayush Bansal  Minh Vo  Yaser Sheikh  Deva Ramanan  Srinivasa Narasimhan
Carnegie Mellon University
{aayushb,mpvo,yaser,deva,srinivas}@cs.cmu.edu
http://www.cs.cmu.edu/~aayushb/Open4D/


Figure 1. We can create virtual cameras that facilitate: (1) freezing the time and exploring views (**red**); (2) freezing the view and moving through time (**green**); and (3) vary both time and view (**blue**).

## Abstract

*We present a data-driven approach for 4D space-time visualization of dynamic events from videos captured by hand-held multiple cameras. Key to our approach is the use of self-supervised neural networks specific to the scene to compose static and dynamic aspects of an event. Though captured from discrete viewpoints, this model enables us to move around the space-time of the event continuously. This model allows us to create virtual cameras that facilitate: (1) freezing the time and exploring views; (2) freezing a view and moving through time; and (3) simultaneously changing both time and view. We can also edit the videos and reveal occluded objects for a given view if it is visible in any of the other views. We validate our approach on challenging in-the-wild events captured using up to 15 mobile cameras.*

## 1. Introduction

Imagine going back in time and revisiting crucial moments of your lives, such as your wedding ceremony, your graduation ceremony, or the first birthday of your child, immersively from any viewpoint. The prospect of building such a *virtual time machine* [40] have become increasingly realizable with the advent of affordable and high-quality smartphone cameras producing extensive collections of social video data. Unfortunately, people do not benefit from this broader set of captures of their social events. When it comes to look back, we are likely to only look at one video or two even when hundreds were captured. We present a data-driven approach that leverages all perspectives to enable a more complete exploration of the event. With our approach, the benefits from each extra perspective that is captured leads to a more complete experience. We seek to automatically organize the disparate visual data into a comprehensive four-dimensional environment (3D space and time). The complete control of spatiotemporal aspects not only enables us to see a dynamic event from any perspective but also allows geometrically consistent content editing. This functionality unlocks many potential applications in the movie industry and consumer devices, especially as virtual reality headsets are becoming popular by the day. Figure 1 show examples of virtual camera views synthesized using our approach for an event captured from multi-view videos.

Prior work on virtualized reality [28, 30, 32] has primarily been restricted to studio setups with tens or even hundreds of synchronized cameras. Four hundred hours of video data is uploaded on YouTube every minute. This feat has become possible because of the commercial success of high qual-


Figure 2. **Comparison to existing work:** Given a dynamic event captured using 10 phones, we **freeze time and explore views** for two time instances. We use a standard Structure-from-Motion (SfM) [42, 43] to reconstruct the camera trajectory. As shown in first-column, SfM treats dynamic information as outliers for rigid reconstruction. We use additional cues such as 2D keypoints [4], statistical human body model [36], and human association [51] along-with the outputs of SfM to generate dynamic information for these two time instances (Frame-450 and Frame-1200 in second and third columns respectively). We call this **SfM+humans**. These three outputs lack *realism*. Additionally, the reconstruction fails for non-Lambertian surfaces (see glass windows), non-textured regions (see umbrellas), and shadows (around humans). Our approach, on the other hand, can densely synthesize the various static and dynamic components, as shown in fourth and fifth columns for the same moments.

ity hand-held cameras such the iPhones or GoPros. Many public events are easily captured from multiple perspectives by different people. Despite this new form of big visual data, reconstructing and rendering the dynamic aspects have mostly been limited to studios and not for in-the-wild captures with hand-held cameras. Currently, there exists no method for fusing the information from multiple cameras into a single comprehensive model that could facilitate content sharing. This gap is largely because the mathematics of dynamic 3D reconstruction [20] is not well-posed. The segmentation of objects [19] are far from being consistently recovered to do 3D reconstruction [56]. Large scale analytics of internet images exist for static scenes [24, 42, 43, 46] alone, and ignores the interesting dynamic events (as shown in Figure 2-first-column).

We pose the problem of 4D visualization from in-the-wild captures within an image-based rendering paradigm utilizing large capacity parametric models. The parametric models based on convolutional neural nets (CNNs) can circumvent the requirement of explicitly computing a comprehensive model [2, 5] for modeling and fusing static and dynamic scene components. Key to our approach is the use of self-supervised CNNs specific to the scene to compose static and dynamic parts of the event. This data-driven model enables us to extract the nuances and details in a dynamic event. We work with in-the-wild dynamic events captured from multiple mobile phone cameras. These multiple views have arbitrary baselines and unconstrained camera poses.

Despite impressive progress with CNN-based scene reconstruction [53, 26, 33, 52], noticeable holes and artifacts are often visible, especially for large texture-less regions or non-Lambertian surfaces. We accumulate spatiotemporal information available from multiple videos to capture content that is not visible at a particular time instant. This accumulation helps us to capture even the large non-textured regions (umbrellas in Figure 2) or non-Lambertian surfaces (glass windows in Figure 2). Finally, a complete control of static and dynamic components of a scene, and viewpoint and time enables user-driven content editing in the videos. In public events, one often encounters random movement obstructing the cameras to capture an event. Traditionally nothing can be done about such spurious content in captured data. The complete 4D control in our system enables the user to remove unwanted occluders and obtain a clearer view of the actual event using multi-view information.

## 2. Related Work

There is a long history of 4D capture systems [30] to experience immersive virtualized reality [14], especially being able to see from any viewpoint that a viewer wants irrespective of the physical capture systems.

**4D Capture in Studios:** The ability to capture depth maps from a small baseline stereo pair via 3D geometry techniques [20] led to the development of video-rate stereo machines [32] mounting six cameras with small baselines. This ability to capture dense depth maps motivated a generation of researchers to develop close studios [28, 31, 39, 58] that can precisely capture the dynamic events happening within it.


Figure 3. **Overview:** We pose the problem of 4D visualization of dynamic events captured from multiple cameras as a data-driven composition of static background (**top**) and instantaneous foreground (**middle**) to generate the final output (**bottom**). Importantly, the data-driven composition enables us to capture certain aspects that may otherwise be missing in the inputs, e.g., parts of the human body are missing in the first and third column, and parts of background are missing in second row.

A crucial requirement in these studios is the use of synchronized video cameras [31]. This line of research is restricted to a few places in the world with access to proper studios and camera systems.

**Beyond Studios:** The onset of mobile phones have revolutionized the capture scenario. Each one of us possess high-definition smartphone cameras. Usually, there are more cameras at a place than there are people around. Many public events are captured by different people from various perspectives. This feat motivated researchers to use in-the-wild data for 3D reconstruction [23, 46] and 4D visualization [2, 5]. A hybrid of geometry [20] and image-based rendering [44] approaches have been used to reconstruct 3D scenes from pictures [7]. Photo tourism [46] and the works following it [1, 15, 16, 24, 45] use internet-scale images to reconstruct architectural sites. These approaches have led to the development of immersive 3D visualization of static scenes.

The work on 3D reconstruction treats dynamic information as outliers and reconstructs the static components alone. Additional cues such as visual hulls [13, 18, 37], or 3D body scans [5, 6], or combination of both [3, 49] are used to capture dynamic aspects (esp. human performances) from multi-view videos. Hasler et al. [21] use markerless method by combining pose estimation and segmentation. Vedula et al. [48] compute scene shape and scene flow for 4D modeling. Ballan et al. [2] model foreground subjects as video-sprites on billboards. However, these methods assume a single actor in multi-view videos. Recent approaches [10, 50] are not restricted by this assumption but does sparse reconstruction.

**CNN-based Image Synthesis:** Data-driven approaches [9, 17, 27, 54] using convolutional neural networks [35] have led to impressive results in image synthesis. These results inspired a large body of work [11, 12, 29, 38, 47, 57] on continuous view synthesis for small baseline shifts. Hedman et al. [22] extended this line of work to free-viewpoint capture. However, these methods are currently applicable to static scenes only. We combine the insights from CNN-based image synthesis and earlier work on 4D visualization to build a data-driven 4D Browsing Engine that makes minimal assumption about the content of multi-view videos.

## 3. 4D Browsing Engine

We are given $N$ camera views with extrinsic parameters $\{C_1, C_2, ..., C_N\}$, and intrinsic parameters $\{M_1, M_2, ..., M_N\}$. Our goal is to generate virtual camera view $C$ that does not exist in any of these $N$ cameras.

In this work, we accumulate the long-term multiview spatiotemporal information to densely reconstruct static background, and combine it with instantaneous information. We pose this problem as a self-supervised composition of static background and instantaneous foreground. Figure 3 shows an overview of our approach via a virtual camera that freezes time and explores views. We begin by describing the overall fusion architecture in Section 3.1, then describe the modules for computing the foreground and background components in Section 3.2, and finally discuss the model in Section 3.3.

### 3.1. Self-Supervised Composition

We use a data-driven approach to learn the fusion of a static background, $B$, and a dynamic foreground, $F$, to generate the required target view for given camera parameters. Since there exists no ground truth or manual annotations, we train a convolutional neural network (CNN) in a self-supervised manner by reconstructing a known held-out camera view, $C$, from the remaining $N-1$ views, thereby learning a mapping $G : (B, F) \rightarrow C$. We use three losses to learn this mapping: (1) Reconstruction loss; (2) Adversarial loss; and (3) Frequency loss.

**Reconstruction Loss:** We use standard $l_1$ reconstruction loss to minimize reconstruction error on the content with paired data samples $\{((b_i, f_i), c_i)\}$ where $b_i \in B$, $f_i \in F$, and $c_i \in C$:

$$\min_G L_r = \sum_i ||c_i - G(b_i, f_i)||_1 \tag{1}$$

**Adversarial Loss:** Recent work [17] has shown that learned mapping can be improved by tuning it with a discriminator $D$ that is adversarially trained to distinguish between real samples of $c_i$ from generated samples $G(b_i, f_i)$:

$$\min_G \max_D L_{adv}(G, D) = \sum_i \log D(c_i) + \sum_i \log(1 - D(G(b_i, f_i))) \tag{2}$$

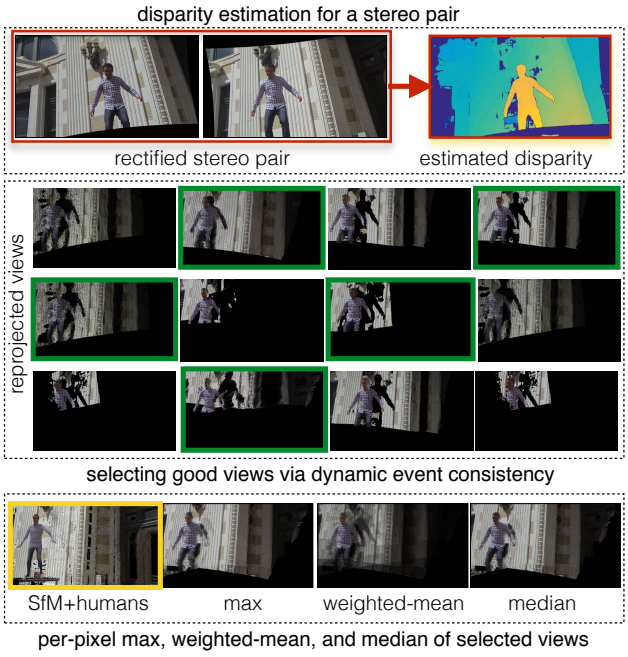


Figure 4. **Instantaneous Foreground Estimation:** We begin with estimating disparity for a stereo pair using an off-the-shelf disparity estimation approach [52]. We use $\binom{N}{2}$ stereo pairs, and reproject them to the target view using standard 3D geometry [20]. We use a dynamic event consistency to select appropriate reprojected views from $\binom{N}{2}$ views (marked with green in **middle**). The dynamic event consistency is computed using the output of SfM+humans (marked with yellow in bottom-row). Finally, we compute a per-pixel max, weighted-mean, and median of selected views. Collectively these represent instantaneous foreground information along with SfM+humans (shown in the **bottom**-row).

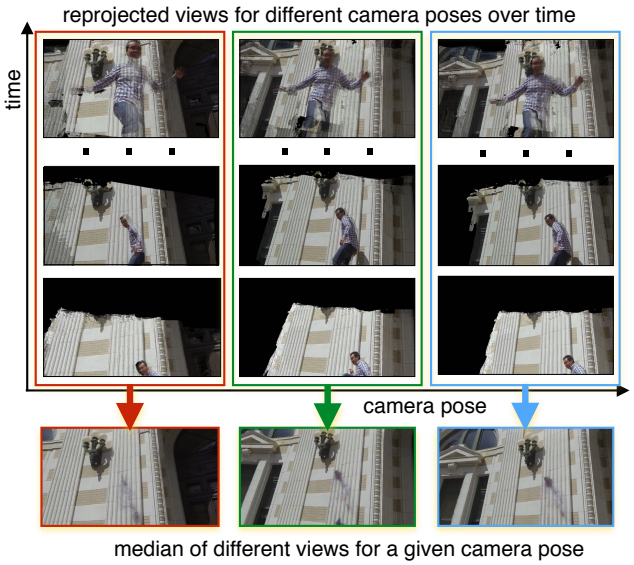


Figure 5. **Static Background Estimation:** We generate images for the target camera pose for all time. A per-pixel median of the images over a large temporal window for the target camera pose filters the dynamic components. We show estimated background images in the **bottom**-row for the three camera poses.

**Frequency Loss:** We enforce a frequency-based loss function via Fast-Fourier Transform to learn appropriate frequency content and avoid generating spurious high-frequencies when ambiguities arise (inconsistent foreground and background inputs):

$$\min_G L_{fr} = \sum_i ||\mathscr{F}(c_i) - \mathscr{F}(G(b_i, f_i))||_1 \qquad (3)$$

where $\mathscr{F}$ is fast-Fourier transform. The overall optimization combines Eq. 1, Eq. 2, and Eq. 3:

$$L = \lambda_r L_r + \lambda_{adv} L_{adv} + \lambda_{fr} L_{fr}$$

where, $\lambda_r = \lambda_{fr} = 100$, and $\lambda_{adv} = 1$. Explicitly using background and foreground for target view makes the model independent of explicit camera parameters.

### 3.2. Intermediate Data Generation

We now describe our approach to estimate information about dynamic foreground and static background that are used as an input to the neural network (Section 3.3). We start with pre-processing of multiple camera views, and then use it to compute background and foreground information.

**Temporal Alignment & Correspondences:** We establish the frame-level temporal alignment for the all cameras using spatiotemporal bundle adjustment [50]. We estimate pixel-level correspondences between a stereo pair using an off-the-shelf disparity estimation [52]. While these correspondences can be noisy, multiple views constraints a better selection of points across $\binom{N}{2}$ stereo pairs.

**Instantaneous Foreground Estimation:** We build foreground estimates at a given time using stereo pairs. We use estimated disparity [52] to warp to the target view. Figure 4 shows different reprojected views from various stereo pairs. Since we have no control over camera placements, the disparity from various $\binom{N}{2}$ stereo pairs are often noisy and cannot be naively used to synthesize the target view in all conditions. Sparse cameras, large stereo baseline, bad stereo-pairs, or errors in camera-poses may result in misaligned frames (shown in Figure-4-middle) for the target camera pose. Therefore we enforce dynamic event consistency via 3D reconstruction to select five best reprojected views to compose instantaneous foreground information.

**Dynamic Event Consistency:** We use a previous approach to perform long-term 3D human tracking across multiple views [51]. The 3D tracking provides a rough 3D estimate of humans from different views. Collectively, with the output of 3D background reconstruction from SfM and MVS [42, 43], we call this **SfM+humans**. While not a *realistic* and *precise*

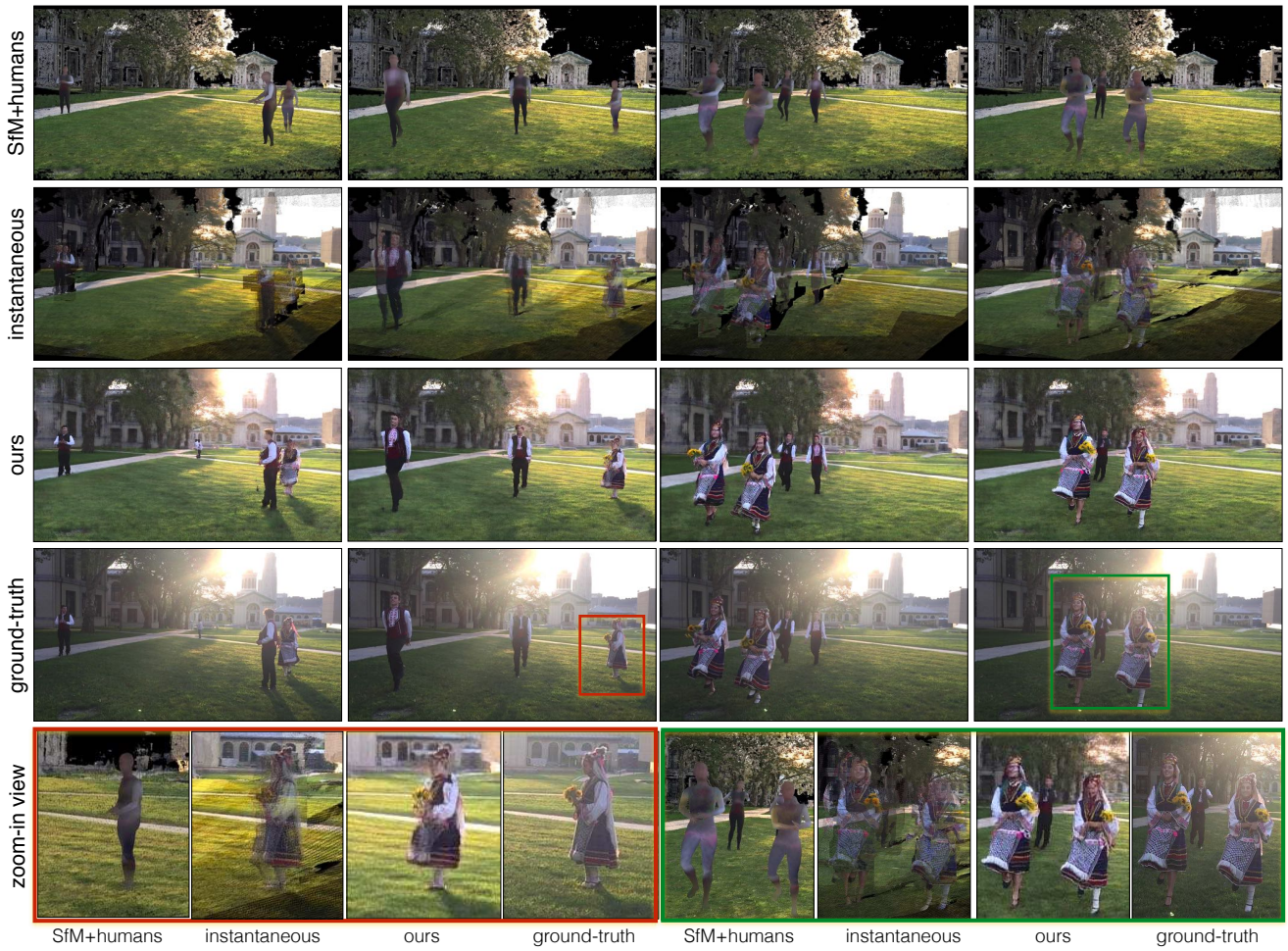


Figure 6. **Freeze a view and Move in time:** We show 4 frames from a 3 minute-long generated video of a stationary camera. This sequence is captured using 10 phones. The **top**-row shows the output generated using SfM+humans. The **second**-row shows intermediate using instantaneous information. We observe missing foreground and background details in the first and second row. The **third** row shows the output of our approach. Our approach can consistently generate the background and foreground. We also compare our outputs to a ground-truth *held-out* camera sequence in **fourth**-row. Finally, we show a few close-ups in the **last**-row. Our approach not only captures the humans well but also contains detailed information such as *flowing* dresses and flowers in it. We are, however, not able to capture the sun's glare at this location as we compose output from views at other locations and do not explicitly parameterize illumination.

output by itself, such 3D reconstruction is sufficient to rank the various stereo pairs that are required to generate a target camera view. This is the main purpose of SfM+Human. We compute the distance between various reprojections and SfM+humans. This distance is computed using the Conv-5 features of an ImageNet [8] pre-trained AlexNet model [34]. We use top-5 scoring views for composing an instantaneous foreground image. As shown in Figure 4, we find good stereo-pairs using SfM+humans (marked yellow). We compute a per-pixel max, weighted-mean, and median using the top-5 ranked stereo pairs (marked green). These images, along with SfM+humans, collectively represent instantaneous information (Figure 4-bottom).

**Static Background Estimation:** We accumulate long-term spatiotemporal information to compute static background for a target camera view. The intrinsic and extrinsic parameters from $N$ physical cameras enable us to create the views over a large temporal window of $[0, t]$ for a target camera position. Figure 5 shows creation of virtual cameras for various poses and time instants. We estimate a static background by computing a median of different views for a given camera pose. Computing the median over large temporal window for a given camera position enables us to capture non-textured and non-Lambertian stationary surfaces in a scene (see Figure 5).

### 3.3. Stacked Multi-Stage CNN

We now describe the neural network architecture that composes the static background and instantaneous fore-


Figure 7. **Many people and their unrestricted movement:** We captured a Jiu-Jitsu retreat event that was witnessed by more than 30 people. We show 4 frames of two virtual cameras (**freeze a view and move in time**) and contrast it with the held-out camera sequences (ground truth). We also show close-ups at various locations and compare it with the ground-truth. We capture various nuance despite people in different clothing, poses, and involved in unchoreographed activities.

ground. We use a modified U-Net [41] architecture that inputs background and foreground information, and outputs the image. Most consumer phones enable to capture $1080p$ videos at 60fps. Training a neural network combining the various background and foreground information with hi-res images require the use of high capacity models. These models need prohibitive memory and hence we use a stacked multi-stage CNN for an effective composition. We use a high capacity model for low-res image generation that learns overall structure, and improve the resolution with multiple stages. We train three models for three different resolutions, namely: (1) low-res ($270\times480$); (2) mid-res ($540\times960$); and (3) hi-res ($1080\times1960$). These models are trained independently and form multiple stages of our formulation. At test time, we use them sequentially, starting from low-res to mid-res to hi-res outputs. The median channel of foreground information for mid-res model is replaced by a $2\times$ upsampled output of low-res model. Similarly, the median channel of foreground information for hi-res model is replaced by a $2\times$ upsampled output of mid-res model. We artificially create

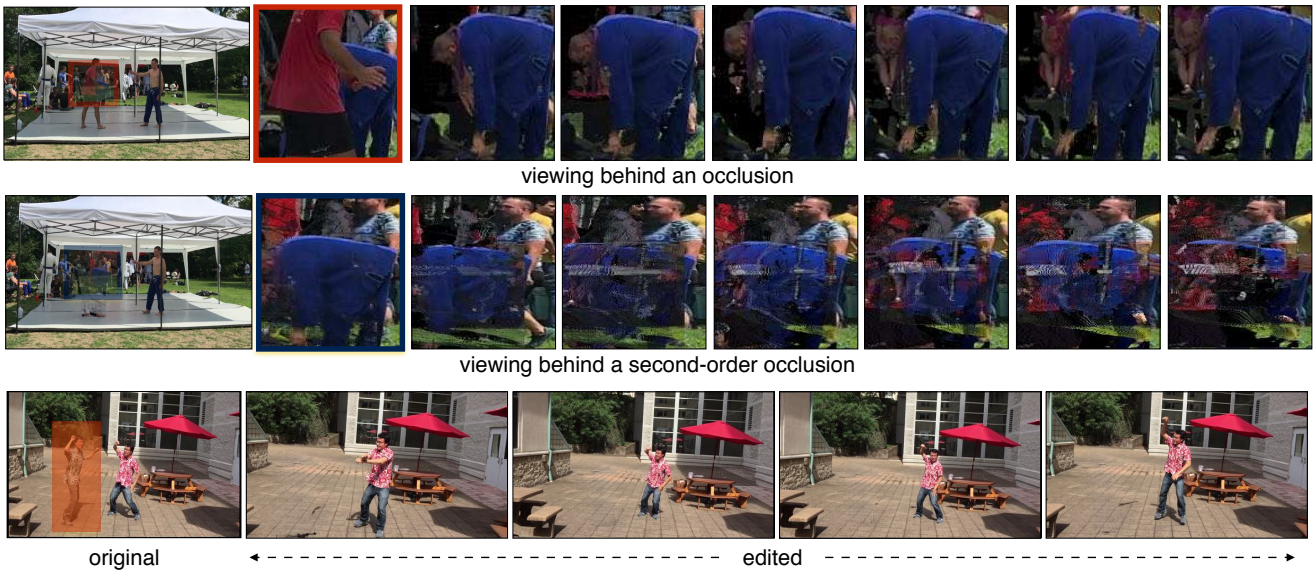


Figure 8. **User-Controlled Manipulation:** We show two examples of user-controlled manipulation and editing in videos. In the **top**-row, a user selects a mask to see the occluded blue-shirt person (behind red shirt person). There is no way we can infer this information from a single-view. However, multi view information allows us to not only see the occluded human but also gives a sense of activity he is doing. We show frames from 2 seconds of video. In the **middle**-row, we want to see the part of scene behind the blue-shirt person who is *disoccluded* above. This is an example of a seeing behind a second-order occlusion. While not as sharp as first-order occlusion result, we can still see green grass and white bench in the background with a person moving. This particular scenario is not only challenging due to second-order occlusion but also because of larger distance from cameras. In the **bottom**-row, a user can remove the foreground person by marking on a single frame in video. Our system associates this mask to all the frames in video, and edit it to show background in place of human. We show frames of edited video (20 seconds long).

a low-resolution median foreground image (down-sampled by a factor of 2) during training to effectively utilize the modifications to mid-res and hi-res models at test time. We provide more details for stacked multi-stage composition on our project page.

### 3.4. User-Controlled Manipulation

We have complete control of the 3D space and time information of the event. This 4D control allows us to browse the dynamic events. A user can see behind the occlusions, edit, add, or remove objects. To accomplish this, a user only needs to mark the required portion in a video. Our approach automatically edits the content, i.e., update the background and foreground, via multi-view information — the modified inputs to stacked multi-stage composition results in desirable outputs. Importantly, marking on a single frame in the video is sufficient, as we can effortlessly propagate the mask to the rest of the video (4D control of foreground). We show two examples of user-controlled manipulation in Figure 8. In the first example, we enable a user to see occluded person without changing the view. Our system takes input of mask from the user, and *disocclude* the blue-shirt person (Figure 8-top-row). We also explore viewing behind a second-order occlusion. Figure 8-middle shows a very challenging example of viewing behind the blue-shirt person. Despite farther away from the camera, we see grass, white table, and a person moving in the output. Finally, we show an example of editing where a user can mark region in a frame of video (Figure 8-bottom-row). Our system generates full video sequence without the masked person.

## 4. Experiments

**Datasets:** We collected a large number of highly diverse sequences of unrestricted dynamic events having a wide variety of human motion, human-human interaction, human-object interaction, clothing, both indoor and outdoor, under varying environmental and illumination conditions. These sequences are captured using upto 15 mobile phones. We refer the reader to our project page for all the results. Here we describe a few prominent sequences used for evaluation.

**Western Folk Dance:** We captured sequences of western folk dance performances. Figure 6 shows the example of one of the sequences from this capture. This sequence is challenging due to *flowing* dresses worn by performers, self-occlusions, and illumination conditions. Such a sequence paves the path for explicit parametrization of illumination condition in 4D modeling.

**Jiu-Jitsu Retreat:** Jiu-Jitsu is a type of Brazilian Martial

| Approach | M.S.E | PSNR | SSIM | LPIPS [55] | FID [25] |
|---|---|---|---|---|---|
| N.N | 1595.78 ±294.53 | 15.50 ±2.72 | 0.476 ±0.086 | 0.401 ±0.027 | - |
| SfM + Humans | 6494.47 ±1721.17 | 9.66 ±1.88 | 0.438 ±0.079 | 0.422 ±0.022 | 184.629 |
| Inst. | 2886.11 ±1654.34 | 13.57 ±3.23 | 0.538 ±0.113 | 0.391 ±0.054 | 122.31 |
| Ours | **591.92 ±286.86** | **20.06 ±3.95** | **0.689 ±0.130** | **0.222 ±0.025** | **47.610** |

Table 1. **Comparison:** We contrast our approach with: (1). a simple nearest neighbor (**N.N.**) baseline; (2). reconstructed outputs of **SfM+humans**; and finally (3). median-channel of instantaneous dynamic information (**Inst**). We use various evaluation crietria to study our approach in comparisons with these three methods: (1). **M.S.E**: We compute a mean-squared error of the generated camera sequences using held-out camera sequences.; (2). **PSNR**: We compute a peak signal-to-noise ratio of the generated sequences against the held out sequences; (3). **SSIM**: We also compute a SSIM in similar manner.; (4). We also use LPIPS [55] to study structural similarity and to avoid any biases due to MSE, PSNR, and SSIM. Lower it is, better it is. Note that all the above four criteria are computed using held-out camera sequences; and finally (5) we compute a **FID**-score [25] to study the quality of generations when a ground-truth is not available for comparisons. Lower it is, better it is.

Art. We captured sequences of this sporting event during a summer retreat of the Jiu-Jitsu group. This sequence is an extreme example of unchoreographed dynamic motion from more than 30 people who participated in it. Figure 1 and Figure 7 show examples from this capture.

**Performance Dance:** We captured many short performance dances including Ballet, Tango, and renactments of plays. The illumination, clothing, and motions change drastically in these sequences.

**Sequences from Prior Work:** We also used sequences from Vo et al. [50] to properly compare our results with their 3D reconstruction (SfM+humans). Figure 2 and Figure 3 shows the results of freezing the time and exploring the views for these sequences.

**Evaluation:** We use a mean-squared error (MSE), PSNR, SSIM, and LPIPS [55] to study the quality of virtual camera views created using our approach. **MSE**: Lower is better. **PSNR**: Higher is better. **SSIM**: Higher is better. **LPIPS**: Lower is better. We use held-out cameras for proper evaluation. We also compute a **FID** score [25], lower the better, to study the quality of sequences where we do not have any ground truth (e.g., freezing the time and exploring views). This criterion contrast the distribution of virtual cameras against the physical cameras.

**Baselines:** To the best of our knowledge, there does not exist a work that has demonstrated dense 4D visualization for in-the-wild dynamic events captured from unconstrained multi-view videos. We, however, study the performance of our approach with: (1) a simple nearest neighbor baseline **N.N.**: We find nearest neighbors of generated sequences using conv-5 features of an ImageNet pre-trained AlexNet model. This feature space helps in finding the images closer in structure.; (2) **SfM+humans**: We use work from Vo et al [50, 51] for these results.; and finally (3) we contrast it with median channel of instantenous image (**Inst**).

Table 1 contrasts our approach with various baselines on held-out cameras for different sequences. In total, we generated data for 12 minutes long sequences for evaluation against held-out sequences, and another 12 minutes of random movements. We observe significantly better outputs under all the criteria. We provide more qualitative and quantitative ablation studies on our project page.

## 5. Discussion & Future Work

The world is our studio. The ability to do 4D visualization of dynamic events captured from unconstrained multi-view videos opens up avenue for future research to capture events with a combination of drones, robots, and hand-held cameras. The use of self-supervised scene-specific CNNs allows one to browse the 4D space-time of dynamic events captured from unconstrained multi-view videos. We extensively captured various in-the-wild events to study this problem. We show different qualitative and quantitative analysis in our study. A real-time user guided system that allows a user to upload videos and browse will enable a better understanding of 4D visualization systems. The proposed formulation and the captured sequences, however, open a number of opportunities for future research such as incorporating illumination and shadows in 4D spatiotemporal representation, and modeling low-level high frequency details. One drawback of our method is that the video streams are treated as perfectly synchronized. This introduces motion artifacts for fast actions [50]. Future work will incorporate sub-frame modeling between different video streams in depth estimation and view synthesis modules for more appealing 4D slow motion browsing.

**Acknowledgements:** We are extremely grateful to Bojan Vrcelj for helping us shape the project. We are also thankful to Gengshan Yang for his help with the disparity estimation code and many other friends for their patience in collecting the various sequences. We list them on our project page. This work is supported by the Qualcomm Innovation Fellowship.

## References


[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. In *ICCV*, 2009. 3

[2] Luca Ballan, Gabriel J. Brostow, Jens Puwein, and Marc Pollefeys. Unstructured video-based rendering: Interactive exploration of casually captured videos. *ACM Trans. Graph.*, 2010. 2, 3

[3] Luca Ballan and Guido Maria Cortelazzo. Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes. 3DPVT, 2008. 3

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018. 2

[5] Joel Carranza, Christian Theobalt, Marcus A. Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. *ACM Trans. Graph.*, 2003. 2, 3

[6] Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. In *ACM SIGGRAPH*. 2008. 3

[7] Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *ACM Trans. Graph.* ACM, 1996. 3

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 5

[9] Emily L Denton, Soumith Chintala, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *NeurIPS*, 2015. 3

[10] N. Dinesh Reddy, Minh Vo, and Srinivasa G. Narasimhan. Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicles. In *CVPR*, 2018. 3

[11] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *CVPR*, 2019. 3

[12] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *CVPR*, 2016. 3

[13] J-S Franco and Edmond Boyer. Fusion of multiview silhouette cues using a space occupancy grid. In *ICCV*, 2005. 3

[14] H. Fuchs, G. Bishop, K. Arthur, L. McMillan, R. Bajcsy, S. Lee, H. Farid, and Takeo Kanade. Virtual space teleconferencing using a sea of cameras. In *Proc. First International Conference on Medical Robotics and Computer Assisted Surgery*, 1994. 2

[15] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. Towards internet-scale multi-view stereo. In *CVPR*, 2010. 3

[16] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *TPAMI*, 2009. 3

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 3

[18] Jean-Yves Guillemaut, Joe Kilner, and Adrian Hilton. Robust graph-cut scene segmentation and reconstruction for free-viewpoint video of complex dynamic scenes. In *ICCV*, 2009. 3

[19] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 2

[20] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2, 3, 4

[21] Nils Hasler, Bodo Rosenhahn, Thorsten Thormahlen, Michael Wand, Jürgen Gall, and Hans-Peter Seidel. Markerless motion capture with unsynchronized moving cameras. In *CVPR*, 2009. 3

[22] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Trans. Graph.*, 2018. 3

[23] Jared Heinly. *Toward Efficient and Robust Large-Scale Structure-from-Motion Systems*. PhD thesis, The University of North Carolina at Chapel Hill, 2015. 3

[24] Jared Heinly, Johannes Lutz Schönberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the World* in Six Days *(As Captured by the Yahoo 100 Million Image Dataset). In *CVPR*, 2015. 2, 3

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 8

[26] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. *CVPR*, 2018. 2

[27] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 3

[28] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE TPAMI*, 2017. 1, 2

[29] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Trans. Graph.*, 2016. 3

[30] Takeo Kanade and PJ Narayanan. Historical perspectives on 4d virtualized reality. In *CVPR Workshops*, 2006. 1, 2

[31] Takeo Kanade, Peter Rander, and PJ Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE multimedia*, 1997. 2, 3

[32] T. Kanade, A. Yoshida, K. Oda, H. Kano, and M. Tanaka. A stereo machine for video-rate dense depth mapping and its new applications. In *CVPR*, 1996. 1, 2

[33] Tejas Khot, Shubham Agrawal, Shubham Tulsiani, Christoph Mertz, Simon Lucey, and Martial Hebert. Learning unsupervised multi-view stereopsis via robust photometric consistency. *arXiv preprint arXiv:1905.02706*, 2019. 2

[34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 5

[35] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015. 3

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 2015. 2

[37] Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J Gortler, and Leonard McMillan. Image-based visual hulls. In *ACM Trans. Graph.*, 2000. 3

[38] Moustafa Meshry, Dan B. Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. In *CVPR*, 2019. 3

[39] Martin Oswald and Daniel Cremers. A convex relaxation approach to space time multi-view 3d reconstruction. In *ICCVW*, 2013. 2

[40] Raj Reddy. *Teleportation, Time Travel, and Immortality*. Springer New York, 1999. 1

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015. 6

[42] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2, 4

[43] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 2, 4

[44] Harry Shum and Sing Bing Kang. Review of image-based rendering techniques. In *Visual Communications and Image Processing*, 2000. 3

[45] Sudipta N Sinha, Drew Steedly, Richard Szeliski, Maneesh Agrawala, and Marc Pollefeys. Interactive 3d architectural modeling from unordered photo collections. In *ACM Trans. Graph.* ACM, 2008. 3

[46] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3d. *ACM Trans. Graph.*, 2006. 2, 3

[47] Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *CVPR*, 2019. 3

[48] Sundar Vedula, Simon Baker, and Takeo Kanade. Image-based spatio-temporal modeling and view interpolation of dynamic events. *ACM Trans. Graph.*, 2005. 3

[49] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *ACM SIGGRAPH 2008*. 2008. 3

[50] Minh Vo, Srinivasa G. Narasimhan, and Yaser Sheikh. Spatiotemporal bundle adjustment for dynamic 3d reconstruction. In *CVPR*, 2016. 3, 4, 8

[51] Minh Vo, Ersin Yumer, Kalyan Sunkavalli, Sunil Hadap, Yaser Sheikh, and Srinivasa Narasimhan. Automatic adaptation of person association for multiview tracking in group activities. *IEEE TPAMI*, 2020. 2, 4, 8

[52] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *CVPR*, 2019. 2, 4

[53] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *CVPR*, 2018. 2

[54] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 3

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 8

[56] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *TPAMI*, 1999. 2

[57] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 2018. 3

[58] C. Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *ACM Trans. Graph.*, 2004. 2