Novel View Synthesis of Dynamic Scenes with Globally Coherent Depths from a Monocular Camera

Jae Shin Yoon[†] Kihwan Kim[♯] Orazio Gallo[♯] Hyun Soo Park[†] Jan Kautz[♯]
[†]University of Minnesota [♯]NVIDIA

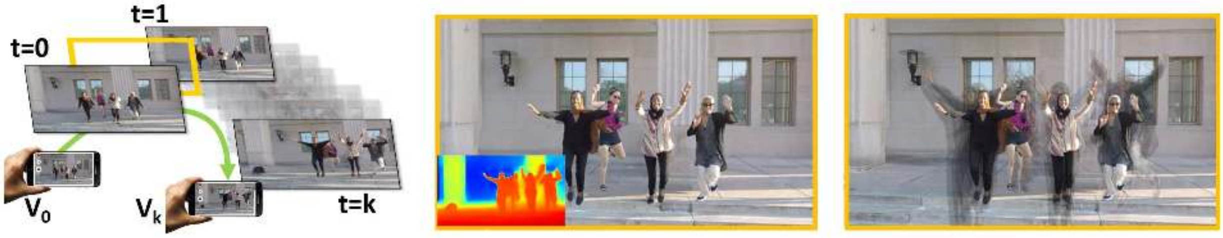

Figure 1: **Dynamic Scene View Synthesis:** (Left) A dynamic scene is captured from a monocular camera from the locations $\mathbf{V}_0$ to $\mathbf{V}_k$. Each image captures people jumping at each time step ($\mathbf{t} = 0$ to $\mathbf{t} = k$). (Middle) A novel view from an arbitrary location between $\mathbf{V}_0$ and $\mathbf{V}_1$ (denoted as an orange frame) is synthesized with the dynamic contents observed at the time $\mathbf{t} = k$. The estimated depth at $\mathbf{V}_k$ is shown in the inset. (Right) For the novel view (orange frame), we can also synthesize the dynamic content that appeared across any views in different time (traces of the foreground in each time step are shown). More results are shown in Sec. 5 and the supplementary document and video.

Abstract

This paper presents a new method to synthesize an image from arbitrary views and times given a collection of images of a dynamic scene. A key challenge for the novel view synthesis arises from dynamic scene reconstruction where epipolar geometry does not apply to the local motion of dynamic contents. To address this challenge, we propose to combine the depth from single view (DSV) and the depth from multi-view stereo (DMV), where DSV is complete, i.e., a depth is assigned to every pixel, yet view-variant in its scale, while DMV is view-invariant yet incomplete. Our insight is that although its scale and quality are inconsistent with other views, the depth estimation from a single view can be used to reason about the globally coherent geometry of dynamic contents. We cast this problem as learning to correct the scale of DSV, and to refine each depth with locally consistent motions between views to form a coherent depth estimation. We integrate these tasks into a depth fusion network in a self-supervised fashion. Given the fused depth maps, we synthesize a photorealistic virtual view in a specific location and time with our deep blending network that completes the scene and renders the virtual view. We evaluate our method of depth estimation and view synthesis on diverse real-world dynamic scenes and show the outstanding performance over existing methods.

1. Introduction

Novel view synthesis [8] is one of the core tasks in computer vision and graphics, and has been used for many visual effects and content creation applications such as cinemagraph [4, 26], video stabilization [28, 21], and bullet time visual effect [60]. In this paper, we focus on view synthesis of dynamic scenes observed from a moving monocular camera as shown in Figure 1. Until now, most of existing view synthesis methods are largely limited to static scenes [8, 60, 9, 11, 56, 59] because they commonly rely on geometric assumptions: in principle, dynamic visual content such as people, pets, and vehicles are considered *outliers* despite being a major focus in videography on social media and otherwise.

Our problem shares the challenge of dynamic scene reconstruction: recovering the underlying 3D geometry of dynamic contents from a moving monocular camera is fundamentally ill-posed [33]. We address this challenge by leveraging the following complementary visual and motion cues.
(1) Multi-view images can be combined to reconstruct incomplete yet view-invariant static scene geometry[1], which enables synthesizing a novel view image of static contents in a geometrically consistent way.
(2) Relative depth predicted from a single image provides view-variant [6] yet complete dynamic scene geometry, which allows enforcing locally consistent 3D scene flow for the foreground dynamic contents.

[1] Its fixed scale chosen from SfM pipeline is consistent across different views from initial triangulation [15].

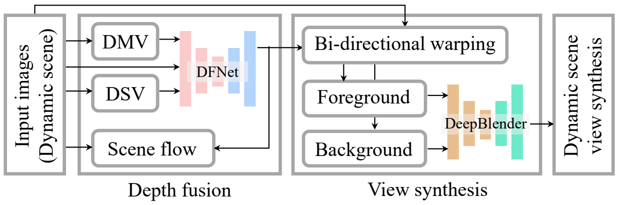

Figure 2: Images of a dynamic scene are used to predict and estimate the depth from single view (DSV) and the depth from multiview stereo (DMV). Our depth fusion network (DFNet) fuses the individual strengths of DSV and DMV (Sec. 3.1) to produce a complete and view-invariant depth by enforcing geometric consistency. The computed depth is used to synthesize a novel view and our DeepBlnder network refines the synthesized image (Sec. 3.2).

We combine these cues by learning a nonlinear scale correction function that can upgrade a time series of single view geometries to form a coherent 4D reconstruction. To disambiguate the geometry of the foreground dynamic contents, we find their simplest motion description in 3D (i.e., slow and smooth motion [49, 40]), which generates minimal stereoscopic disparity when seen by a novel view [5].

We model the scale correction function using a depth fusion network that takes input images, view-variant depth from single view (DSV), and incomplete yet view-invariant depth from a multi-view stereo (DMV) algorithm, and outputs complete and view-invariant depth. The network is self-supervised by three visual signals: (i) the static regions of the DSV must be aligned with a DMV; (ii) the output depth of dynamic regions must be consistent with the relative depth of each DSV; and (iii) the estimated scene flow must be minimal and locally consistent. With the predicted depths that are geometrically consistent across views, we synthesize a novel view using a self-supervised rendering network that produces a photorealistic image in the presence of missing data with adversarial training. An overview of our pipeline is shown in Figure 2.

We show that the novel view synthesis with our depth prediction method is highly effective in generating an unseen image. Further, the rendering network seamlessly blends foreground and background, which outperforms existing synthesis approaches quantitatively and qualitatively.

Our key contributions are as follows:

- A novel depth fusion network that models a scale correction function, which completes the depth maps of view-invariant dynamic scene geometry with locally consistent motions.
- A rendering network that combines foreground and background regions in a photorealistic way using adversarial training.
- A real-world dataset captured by fixed baseline multi-view videos and corresponding benchmark examples for dynamic scene view synthesis.

2. Related Work

For the view synthesis of a dynamic scene from images with baselines, the depth and foreground motion from each view need to be consistent across the views. Here we review view synthesis, depth estimation, and scene reconstruction techniques, and discuss the relations to our method.

Novel View Synthesis The problem of novel view synthesis is strongly tied to multiview 3D reconstruction as it requires transporting pixels across views through the geometry of scenes. For a static scene, multiview geometry [15] applies, which allows triangulating correspondences given the calibrated cameras. This leads to dense reconstruction (multiview stereo [14]), which allows continuous view synthesis using a finite number of images [13]. In principle, the triangulation requires a correspondence from at least two views. This requirement often cannot be met in particular for occluded regions, which results in an incomplete view synthesis. Such issue escalates as the baseline between images increases. Various scene priors have been used to mitigate this issue. A scene that can be expressed as a set of planar surfaces can be reconstructed even with a single view image [45, 27], and a scene with known object categories can be reconstructed with the shape priors [16]. For dynamic scene view synthesis, synchronized multiple cameras are used where the same geometric principle can be applied [60, 29, 18]. Recently, single view depth prediction that is learned from a large image repository is used to complete the scene geometry [37, 25, 23], and even enables to extrapolate the views beyond the range of camera motion [46, 9]. More recent approach, which predicts the depth from single view with human specific priors, realizes the view synthesis of dynamic scene of moving people from a monocular camera [24]. Our approach is inspired by learning based scene completion [12, 37, 59, 24] while applying to a dynamic scene with geometric consistency without any category-specific priors.

Monocular Dynamic Scene Reconstruction Dynamic scene reconstruction using a moving monocular camera, without a prior assumption of scenes, is very challenging and ill-posed, similar to reconstructing a scene with a single view image. For a temporal prior, trajectory triangulation extends the concept of point triangulation by representing scene trajectories using a family of algebraic groups, e.g., line/conic [3], homography tensors on the plane [44, 51], polynomials [19], and discrete cosine transform basis [1, 33]. For spatial prior, a shape can be expressed by a linear combination of compact shape basis vectors [7], which is highly effective to describe constrained deformation such as face. A key challenge is learning shape basis vectors for unknown objects, which requires additional spatial priors such as orthogonality of basis vectors [53], temporal smoothness [48, 32, 49], articulation constraint using joint subspace [54], local piecewise rigid-

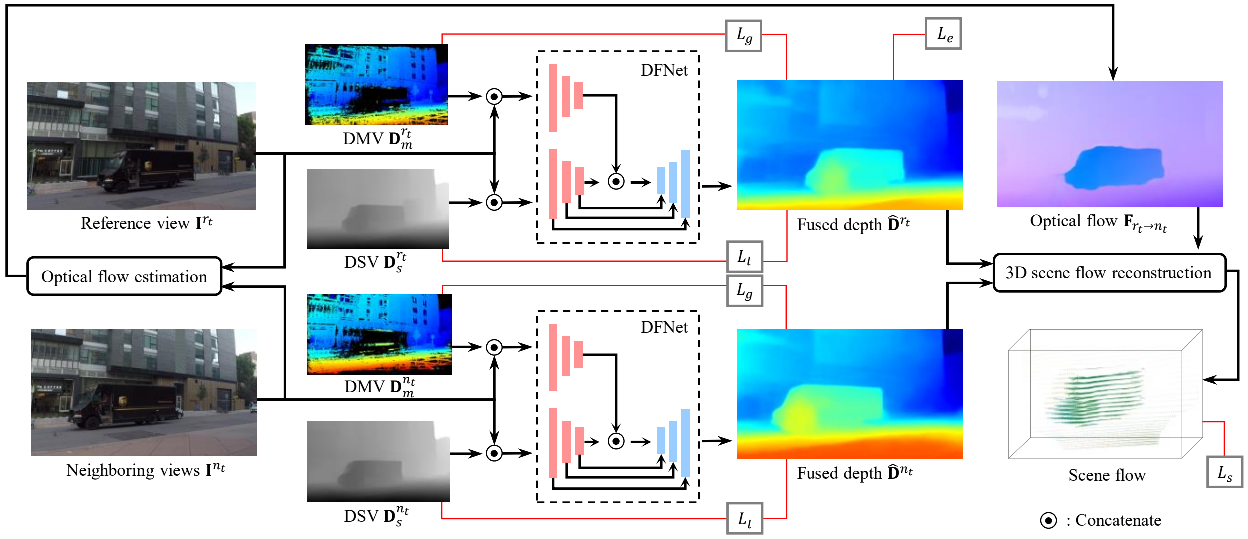
DMV $\mathbf{D}_m^{r_t}$
Reference view $\mathbf{I}^{r_t}$
DSV $\mathbf{D}_s^{r_t}$
Optical flow estimation
DFNet
L_g
L_e
L_l
Fused depth $\hat{\mathbf{D}}^{r_t}$
Optical flow $\mathbf{F}_{r_t \to n_t}$
3D scene flow reconstruction
DMV $\mathbf{D}_m^{n_t}$
Neighboring views $\mathbf{I}^{n_t}$
DSV $\mathbf{D}_s^{n_t}$
DFNet
L_g
L_l
Fused depth $\hat{\mathbf{D}}^{n_t}$
Scene flow
L_s
⊙ : Concatenate

Figure 3: Depth Fusion Network (DFNet) predicts a complete and view-invariant depth map by fusing DSV and DMV with the image. DFNet is self-supervised by minimizing the background depth consistency with DMV (L_g), the relative depth consistency with DSV (L_l), 3D scene flow (L_s), and spatial irregularity (L_e).

ity [10], and learning from training data [20]. For completeness, image regions are reconstructed independently using shape basis [41] or local patches [22], which can be stitched together to form complete scene reconstruction. Further, the spatial and temporal priors can be combined to produce dense correspondences, resulting in complete 4D reconstruction [17]. Humans are a special case of spatial constraints, which allow markerless motion capture from a monocular camera [52, 35, 2]. Unlike the explicit spatial priors, our work makes use of general geometric priors and motion constraint to reconstruct a complete and view-invariant geometry of general dynamic scenes, which allows us to generate realistic spatio-temporal view synthesis.

3. Approach

We cast the novel view synthesis problem as image warping from input source views to a virtual view using underlying 4D reconstruction, i.e.,

$$\mathbf{J}^v(W_{r \to v}(\mathbf{x})) = \mathbf{I}^r(\mathbf{x}), \tag{1}$$

where $\mathbf{J}^v$ is the synthesized image from an arbitrary virtual view v (v can be a source viewpoint), $W_{r \to v}$ is a warping function, and $\mathbf{I}^r$ is the r^{th} source image.

For view synthesis of static scene, the warping function can be described as:

$$\mathbf{y} = W_{r \to v}(\mathbf{x}; \mathbf{D}^r, \Pi^r, \Pi^v), \tag{2}$$

where Π^r and Π^v are the projection matrices at the r^{th} and v^{th} viewpoints. The warping function forms the warped coordinates $\mathbf{y}$ by reconstructing the view-invariant 3D geometry using the depth ($\mathbf{D}^r$) and projection matrix at the r^{th} viewpoint, and projecting onto the v^{th} viewpoint. For instance, this warping function can generate the i^{th} source image from the j^{th} source image, i.e., $\mathbf{I}^i(W_{j \to i}) = \mathbf{I}^j$.

For view synthesis of dynamic scene, the warping function can be generalized to include the time-varying geometry using the depth $\mathbf{D}^{r_t}$, i.e.,

$$\mathbf{y} = W_{r_t \to v}(\mathbf{x}; \mathbf{D}^{r_t}, \Pi^r, \Pi^v), \tag{3}$$

where r_t is the time dependent view index, and t is the time instant. Note that for a moving monocular camera, the view is a function of time. Unlike the static scene warping $W_{r \to v}$ in Eq. (2), we cannot synthesize i^{th} source image from the j^{th} source image because of the time-varying geometry $\mathbf{D}^{r_t}$, i.e., $\mathbf{I}^i(W_{j \to i}) \neq \mathbf{I}^j$.

With these two warping functions, the dynamic scene view synthesis can be expressed as:

$$\mathbf{J} = \phi\left(\left\{\mathbf{J}^v\left(W_{r \to v}\right)\right\}_r, \mathbf{J}^{v,t}\left(W_{r_t \to v}\right); \mathcal{M}^v\right), \tag{4}$$

where $\{\mathbf{J}^v(W_{r \to v})\}_r$ is a set of static scene warping from all source viewpoints, and $\mathbf{J}^{v,t}(W_{r_t \to v,t})$ is the warping of dynamic contents from the source image of the t^{th} time instant. $\mathcal{M}^v$ is the set of the coordinates belonging to dynamic contents. ϕ is the rendering function that refines the warped images to complete the view synthesis.

In Eq. (4), two quantities are unknowns: the depth from each source view $\mathbf{D}^{r_t}$ and the rendering function ϕ. We formulate these two quantities in Sec. 3.1 and Sec. 3.2.

3.1. Globally Coherent Depth from Dynamic Scenes

Our conjecture is that there exists a scale correction function that can upgrade a complete view-variant depth $\mathbf{D}_s^{r_t}$

from the single view prediction (DSV) to the depth of the view-invariant 3D geometry $\widehat{\mathbf{D}}^{r_t}$:

$$\widehat{\mathbf{D}}^{r_t} = \psi(\mathbf{D}_s^{r_t}), \tag{5}$$

where ψ is the scale correction function. Ideally, when a scene is stationary, the upgraded depth is expected to be identical to the depth $\mathbf{D}_m^r$ from view-invariant geometry, e.g., depth from multiview stereo (DMV), with uniform scaling, i.e., $\mathbf{D}_m^r = \psi(\mathbf{D}_s) = \alpha\mathbf{D}_s + \beta$ where α and β are scalar and bias. When a scene is dynamic, the linear regression of such scale and bias is not applicable. We learn a nonlinear scale correction function that possesses the following three properties.

First, for the static scene, the upgraded depth approximates DMV:

$$\mathbf{D}_m^r(\mathbf{x}) \approx \psi\left(\mathbf{D}_s^{r_t}(\mathbf{x})\right) \quad \text{for } \mathbf{x} \notin \mathcal{M}^{r_t}, \tag{6}$$

where $\mathbf{x}$ is the coordinate of pixels belonging to the static background.

Second, for the dynamic contents, the upgraded depth preserves the relative depth from DSV:

$$g\left(\mathbf{D}_s^{r_t}(\mathbf{x})\right) \approx g\left(\psi\left(\mathbf{D}_s^{r_t}(\mathbf{x})\right)\right) \quad \text{for } \mathbf{x} \in \mathcal{M}^{r_t}, \tag{7}$$

where g measures the scale invariant relative gradient of depth, i.e.,

$$g(\mathbf{D}; \mathbf{x}, \Delta\mathbf{x}) = \frac{\mathbf{D}(\mathbf{x}+\Delta\mathbf{x}) - \mathbf{D}(\mathbf{x})}{|\mathbf{D}(\mathbf{x}+\Delta\mathbf{x})| + |\mathbf{D}(\mathbf{x})|}. \tag{8}$$

We use multi-scale neighbors $\mathbf{x}+\Delta\mathbf{x}$ to constrain local and global relative gradients.

Third, 3D scene motion induced by the upgraded depths is smooth and slow [50], i.e., minimal scene flow:

$$\mathbf{p}(\mathbf{x}; \mathbf{D}^{r_t}, \Pi^{r_t}) \approx \mathbf{p}(F_{r_t \to n_t}(\mathbf{x}); \mathbf{D}^{n_t}, \Pi^{n_t}), \tag{9}$$

where $F_{r_t \to n_t}$ is the optical flow from the r_t^{th} to n_t^{th} source images. $\mathbf{p}(\mathbf{x}; \mathbf{D}) \in \mathbb{R}^3$ is the reconstructed point in the world coordinate using the depth $\mathbf{D}$:

$$\mathbf{p}(\mathbf{x}; \mathbf{D}, \Pi) = \psi\left(\mathbf{D}(\mathbf{x})\right)\mathbf{R}^{\mathsf{T}}\mathbf{K}^{-1}\widetilde{\mathbf{x}} + \mathbf{C} \tag{10}$$

where $\widetilde{\mathbf{x}}$ is the homogeneous representation of $\mathbf{x}$, and $\mathbf{R} \in SO(3)$, $\mathbf{C} \in \mathbb{R}^3$, and $\mathbf{K}$ are the camera rotation matrix, camera optical center, and camera intrinsic parameters from the projection matrix Π.

Depth Fusion Network (DFNet) We enable the scale correction function ψ using a depth fusion network that takes as input DSV, DMV, and image $\mathbf{I}^{r_t}$:

$$\widehat{\mathbf{D}}^{r_t} = \psi\left(\mathbf{D}_s^{r_t}, \mathbf{D}_m^{r_t}, \mathbf{I}^{r_t}; \mathbf{w}\right), \tag{11}$$

where the network is parametrized by its weights $\mathbf{w}$. To learn $\mathbf{w}$, we minimize the following loss:

$$L(\mathbf{w}) = L_g + \lambda_l L_l + \lambda_s L_s + \lambda_e L_e, \tag{12}$$

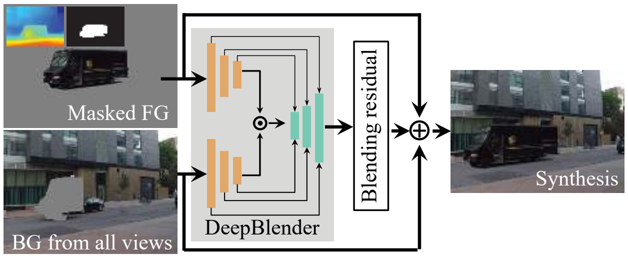

Figure 4: View synthesis pipeline: Given the warped foreground (FG) and background (BG) through the depths and masks, we complete the dynamic scene view synthesis using a rendering network called DeepBlender that predicts the missing region and refines the artifacts.

where λ controls the importance of each loss. L_g measures the difference between DMV and the estimated depth in Eq. (6) for static scene:

$$L_g = \|\widehat{\mathbf{D}}^{r_t}(\mathbf{x}) - \mathbf{D}_m^{r_t}(\mathbf{x})\| \quad \text{for } \mathbf{x} \notin \mathcal{M}^{r_t},$$

L_l compares the scale invariant depth gradient between DSV and the estimated depth in Eq. (7):

$$L_l = \|g(\widehat{\mathbf{D}}^{r_t}(\mathbf{x})) - g(\mathbf{D}_s^{r_t})(\mathbf{x})\| \quad \text{for} \quad \mathbf{x} \in \mathcal{M}^{r_t},$$

and L_s minimize the induced 3D scene motion for entire pixel coordinates in Eq. (9):

$$L_s = \|\mathbf{p}(\mathbf{x}; \mathbf{D}^{r_t}, \Pi^{r_t}) - \mathbf{p}(F_{r_t \to n_t}(\mathbf{x}); \mathbf{D}^{n_t}, \Pi^{n_t})\|.$$

In conjunction with self-supervision, we further minimize the Laplacian of the estimated depth as regularization, i.e.,

$$L_e = \|\nabla^2\widehat{\mathbf{D}}^{r_t}(\mathbf{x})\|^2 + \lambda_f\|\nabla^2\widehat{\mathbf{D}}^{r_t}(\bar{\mathbf{x}})\|^2 \tag{13}$$

where $\mathbf{x} \notin \mathcal{M}^{r_t}$, $\bar{\mathbf{x}} \in \mathcal{M}^{r_t}$, and λ_f balances the spatial smoothness between the static and dynamic regions.

The overview of our self-supervision pipeline and the network architecture are described in Figure 3. DFNet extracts the visual features from DSV and DMV using the same encoder in conjunction with the image. With the visual features, DFNet generates a complte and view invariant depth map that is geometrically consistent. To preserve the local visual features, skip connections between the feature extractor and depth generator are used.

3.2. Dynamic Scene View Synthesis

Given a set of warped static scenes from all source views $\{\mathbf{J}^v\}_r$, we construct a global background $\mathbf{J}_*^v$ based on the baseline between the virtual and source cameras, i.e., assign the pixel value from the warped source view that has the shortest baseline with virtual camera. With $\mathbf{J}_*^v$ and the warped dynamic contents $\mathbf{J}^{v,t}$ from a single time instant, we model the synthesis function ϕ in Eq. (4) as follows:

$$\phi(\mathbf{J}_*^v, \mathbf{J}^{v,t}; \mathcal{M}^v) = \mathbf{J}_*^v(\mathbf{x}) + \mathbf{J}^{v,t}(\mathbf{y}) + \widetilde{\phi}_\theta(\mathbf{J}_*^v, \mathbf{J}^{v,t}), \tag{14}$$

where $\mathbf{x} \notin \mathcal{M}^{v,t}$ and $\mathbf{y} \in \mathcal{M}^{v,t}$. $\widetilde{\phi}_\theta$ is the blending residual that fills the missing regions (unlike a static scene, there exists the regions that are not seen by any source views for a dynamic scene) and refines the synthesized image. We model this blending residual $\widetilde{\phi}_\theta$ using our rendering network.

DeepBlender Network The DeepBlender predicts the blending residual $\widetilde{\phi}_\theta$ from the inputs of a warped dynamic scene $\mathbf{J}^{v,t}$ and a globally modeled static scene $\mathbf{J}^v_*$ as shown in Figure 4. It combines visual features extracted from $\mathbf{J}^{v,t}$ and $\mathbf{J}^v_*$ to form a decoder with skip connections. We learn this rendering function using source images with self-supervision. Each image is segmented into background and foreground with the corresponding foreground mask. We synthetically generate the missing regions near the foreground boundary and image border, and random pixel noises across the scenes. From the foreground and background images with missing regions and pixel noises, the DeepBlender is trained to generate the in-painting residuals. We incorporate an adversarial loss to produce photorealistic image synthesis:

$$L(\mathbf{w}_\theta) = L_{\text{rec}} + \lambda_{\text{adv}} L_{\text{adv}}, \tag{15}$$

where L_{rec} is the reconstruction loss (difference between the estimated blending residual and ground truth), and L_{adv} is the adversarial loss [34]. The overview of our view synthesis pipeline is described in Figure 4.

4. Implementation Details

DFNet is pre-trained on a synthetic dataset [30] (which provides ground-truth optical flow, depth, and foreground mask) for better weight initialization during the self-supervision. To simulate the characteristic of the real data from synthetic, we partially remove the depth around the foreground region and add the depth noise across the scenes with 5% tolerance of the variance at every training iteration. The same self-supervision loss as Eq. 15 is used to pre-train the network. To avoid the network depth scale confusion, we use the normalized inverse depth [23] for both DMV and DSV and recover the scale of the fused depth based on the original scale of DMV. To obtain DSV and DMV, we use existing single view prediction [23] and multiview stereo method [43]. In Eq. (8), we use five multi-scale neighbors, i.e., $\Delta\mathbf{x} = \{1, 2, 4, 8, 16\}$ to consider both local and global regions. We use PWCNet [47] to compute the optical flow in Eq. (9), where the outliers were handled by forward-backward flow consistency. When enforcing the scene flow loss, we use ± 2 neighboring camera views, i.e., $n_t = r_t \pm 2$. We extract the foreground mask using interactive segmentation tools [39]. The foreground masks are manually specified for all baselines in the evaluation, while existing foreground segmentation approaches [38] can be used as a complementary tool as shown in Figure 7.

We also pre-train the DeepBlender using video object segmentation dataset [36]. To create the synthetic residual, we randomly generate the seams and holes around the foreground using mask morphology and superpixel, and remove one side of the image boundary up to 30-pixel thickness. The loss in Eq. 15 is used for pre-training as well. When we warp an image to a virtual view, we check bidirectional warping consistency to prevent the pixel holes. For each image warping, we refine the depth using the bilateral weighted median filters [57]. As shown in Figure 4, we handle the foreground and background separately to prevent the pixel mixing problem around the object boundary.

5. Experiments

We evaluate our method with various dynamic scenes.

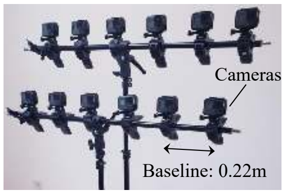

Figure 5: Camera rig.

Dynamic Scene Dataset We collect dynamic scenes using two methods. (1) Moving monocular camera: short-term dynamic events ($\sim$ 5s) are captured by a hand-held monocular moving camera (Samsung Galaxy Note 10) with 60Hz framerate and 1920×1080 resolution. We sub-sample the sequence if the object motion is not salient, and therefore, the degree of the scene motion is significantly larger than that of the camera egomotion where quasi-static dynamic reconstruction does not apply. Four dynamic scenes are captured, which includes human activity, human-object interaction, and animal movement (see the supplementary video). These scenes are used for the qualitative evaluation, where we use half-resolution inputs. (2) Stationary multiview cameras: 8 scenes are captured by a static camera rig with 12 cameras (GoPro Black Edition), where the ground truth of depth estimation and view synthesis are available for the quantitative evaluation. The cameras are located at two levels, and at each level, 6 cameras are evenly distributed with 0.22m baseline as shown in Figure 5. All cameras are manually synchronized. The dataset is categorized into following: (1) Human: a single or multiple people show their dynamic motion, e.g., dynamic facial expression and body motion. (2) Interaction: a person interacts with objects, e.g., umbrella, balloon, and skate. (3) Vehicle: a truck rigidly move from the right side of the road to the left. (4) Stop motion: a doll is sequentially captured in the different location. When testing, we use a set of images sampled from each camera at different time instant to simulate a moving monocular camera. Given the set of collected images, we calibrate the intrinsic and extrinsic parameters of the moving camera using structure-from-motion [42].

Quantitative Evaluation Metric We evaluate the accuracy of depth estimation and view synthesis using the multiview dataset. (1) Depth estimation: given the estimated depth,

F+B / F-only ↘	Jumping	Skating	Truck	DynaFace	Umbrella	Balloon1	Balloon2	Teadybear	Avg.
MVS [43]	0.53 / 2.12	0.29 / 6.81	0.52 / 2.94	0.05 / 0.21	**0.35** / 4.70	0.13 / 1.72	**0.04** / 0.31	**0.06** / 0.92	0.24 / 2.46
RMVSNet [55]	0.61 / 1.55	0.76 / 1.56	0.84 / 2.43	2.24 / 1.57	0.67 / 5.24	0.23 / 1.40	0.13 / 0.38	0.58 / 0.89	0.75 / 1.87
MonoDepth [23]	1.79 / 2.55	1.34 / 2.02	2.62 / 3.86	0.39 / 0.74	2.69 / 4.75	1.07 / 1.88	1.06 / 0.99	0.76 / 0.28	1.46 / 2.13
Sparse2Dense [31]	1.35 / 3.26	1.35 / 10.66	2.15 / 7.60	0.20 / 0.34	1.35 / 6.40	0.53 / 3.03	0.48 / 0.65	0.32 / 0.90	0.96 / 4.10
DFNet-L_g	1.26 / 1.31	0.81 / 0.76	1.60 / 1.24	0.26 / 0.91	2.19 / 1.98	0.93 / 1.36	0.53 / 0.30	1.91 / 0.97	1.18 / 1.10
DFNet-L_l	0.46 / 1.58	0.15 / 1.38	0.62 / 3.34	0.09 / 0.26	0.58 / 3.14	0.15 / 1.57	0.08 / 0.30	0.16 / 0.67	0.28 / 1.53
DFNet-L_e	0.38 / 0.93	0.14 / 0.47	0.52 / 1.09	0.07 / 0.12	0.52 / 2.48	0.15 / 1.20	0.06 / 0.24	0.17 / 0.48	0.26 / 0.87
DFNet-L_s	0.37 / 1.09	0.14 / 0.51	0.53 / 1.11	0.07 / 0.13	0.59 / 2.54	0.16 / 1.18	0.07 / 0.25	0.16 / 0.52	0.26 / 0.91
DFNet	**0.35 / 0.76**	**0.12 / 0.40**	**0.41 / 0.83**	**0.03 / 0.08**	0.37 / **1.90**	**0.12 / 1.11**	0.05 / **0.23**	0.17 / **0.32**	**0.20 / 0.70**

Table 1: Results of quantitative evaluation for the task of depth estimation from dynamic scenes. RMSE in the metric scale is used for evaluation. F and B represent the foreground and background, respectively. The lower is the better.

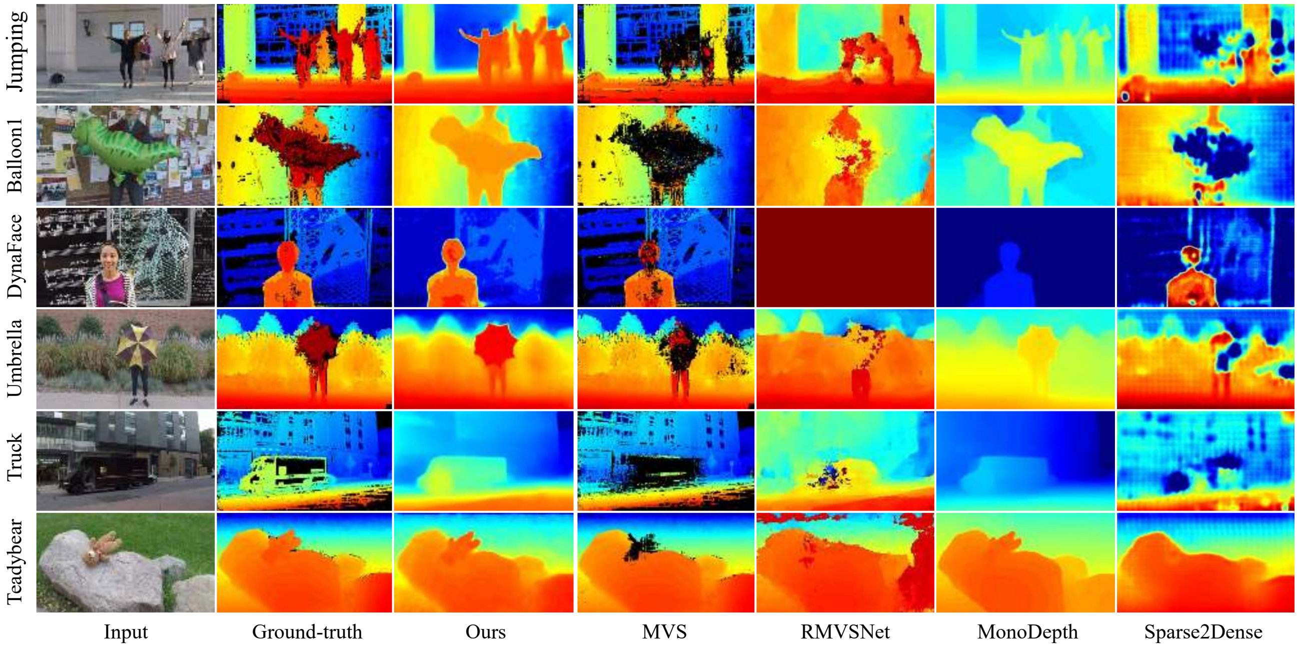

Figure 6: Qualitative comparison of the dynamic scene depth estimation from each method.

we measure root mean square error (RMSE) by comparing to the ground-truth depth computed by multiview stereo. The error is represented in metric scale (m), i.e., the scale of the estimated depth is upgraded to the metric space using the physical length of the camera baseline. We exclude the region that cannot be reconstructed by multiview stereo. (2) View synthesis: we measure the mean of the optical flow magnitude from the ground-truth image to the synthesized one to validate the view invariant property of the depth map. Ideally, it should be close to 0 with the perfect depth map. Additionally, we measure the perceptual similarity [58] (i.e., the distance of VGG features) with the ground-truth to evaluate the visual plausibility of the synthesized view, where its range is normalized into [0, 1] (the lower is the better).

Baselines and Ablation Study We compare our depth estimation and view synthesis methods with a set of baseline approaches. For the depth evaluation, we compare our method with four baselines: 1) Multiview stereo (MVS [43]) assumes that a scene is stationary. For the pixel of which MVS failed to measure the depth, we assign the average of valid depth. 2) RMVSNet [55] is a learning based multiview stereo algorithm. 3) MonoDepth [23] predicts the depth from a single view image. As it produces the normalized depth, we re-scale the predicted depth by using the mean and standard deviation from MVS depth. 4) Sparse2Dense [31] completes the depth given an incomplete depth estimation, where we use MVS depth as an input. As this method requires the metric depth, we upgrade the estimated depth to the metric space using the physical length of the camera baseline. In conjunction with comparative evaluations, we conduct an ablation study to validate the choice of losses.

For the view synthesis evaluation, we compare our view warping method (bi-directional 3D warping) with as-similar-as-possible warping [28] which warps an image by estimating grid-wise affine transforms. The correspondences of the warping are computed by projecting the estimated depth, i.e., transporting pixels in a source image to a novel view through the view-invariant depth. In Table 2, we denote bi-directional warping followed by the DeepBlender refinement as B3W, and as-similar-as-possible warping followed by the DeepBlender as ASAPW. Note that the DeepBlender refinement is applied to all methods except for DFNet+B3W-DeepBlender which evaluates the effect of the refinement by eliminating the DeepBlender. On top of the comparison with different warping methods, we also test all possible combination of depth estimation methods with view warping methods as listed in Table 2. It quantifies how the quality of depth maps affect the view synthesis results.

Dynamic Scene Depth Estimation In Table 1, we summarize the accuracy of dynamic scene depth estimation results evaluated on: 1) the entire scene, and 2) the only dynamic contents. For the entire scene, our method shows the best results on average, followed by MVS with 0.04 m accuracy gap. In the sequence of umbrella and teadybear, MVS shows the better accuracy for the entire scene than ours due to the highly occupant background area as shown in Figure 6, i.e., the depth estimation of dynamic contents much less contributes to depth accuracy evaluation than one of the background. From the evaluation on the only dynamic contents, our method (DFNet) also shows the best result with the noticeable accuracy improvement (1.17 m) from the second best method (MonoDepth).

While the relative depth of MonoDepth is well reflective of the ground-truth, its depth range is often biased to a specific range, e.g., the foreground object is located much closer to the background scenes. Sparse2Dense does not fully reconstruct the background depth even with the MVS depth as inputs, and the predicted foreground depth is completely incorrect. It indicates that fusing the individual strength of learning-base and stereo-based geometry is essential to obtain the globally coherent and complete depth map from dynamic scenes. From Figure 6, we can further notice that the learning based multiview stereo (RMVSNet) also fail to model the dynamic foreground geometry. In our experiment, RMVSNet completely fail when the object is too close to the camera.

From the ablation study described in Table 1, L_g is the most critical self-supervision signal as the MVS depth plays the key role to convey the accurate static depth. Those accurate depths play the fiducial point for the other self-supervision signals to predict the depth on the missing area. From DFNet-L_l, we can verify that the single view depth estimation can upgrade the depth accuracy around the dynamic contents by guiding it with accurate relative depths. Although the contribution of L_e and L_s are relatively small than others, it helps to regularize the object scene motion and the spatial smoothness of the foreground depth which are keys to reduce the artifacts of the novel view synthesis.

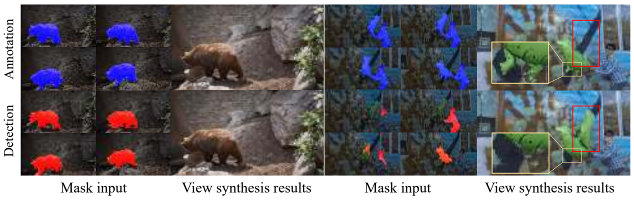

Figure 7: The mask detection with small mistakes (left) does not have a significant impact on the view synthesis results. However, if the mask detection is completely failed (right), it produces artifacts such as object fragmentation (yellow box) or afterimage (red box).

Dynamic Scene Novel View Synthesis Table 2 shows the quantitative evaluation of view synthesis, and the associated qualitative results are shown in Figure 8. From the qualitative results, we can notice that two types of artifacts can be produced depending on the warping methods: B3W produces flying pixel noises, i.e., a pixel is floating due to the warping with incorrect depths, while ASAPW produces image distortion. Such artifacts lead to the increase of the perceptual distance with ground-truths as it captures the structural similarity. On average, our method (DFNet+B3W) shows the smallest perceptual distance (0.15), indicating that the geometry from our depth map is highly preservative of scene structure. The comparison of DFNet+B3W with DFNet+ASAPW demonstrates that, given an accurate depth map, pixel-wise warping (B3W) is the better choice over the grid-wise warping (ASAWP) for view synthesis. From the results of DFNet+B3W-Deepblender, we can observe the large improvement of perceptual similarity compared to the results without DeepBlender, indicating that the refinement step (hole filling and noise reduction) is essential for visual plausibility.

Our method (DFNet+B3W) performs the best even for the flow evaluation (5.3 pixels). MVS+B3W is following our method with the 6.8 pixel errors but it produces a significant pixel noise around the dynamic contents as shown in Figure 8. While MonoDepth+B3W reconstructs visually plausible results in Figure 8, it accompany with large flow errors (10.8 pixels on average), meaning that this result is not geometrically plausible. Note that, the optical flow error of DFNet+B3W-Deepblender is much higher than DFNet+B3W because the flow estimation algorithm [47] shows significant confusion when there are holes around the image boundary and dynamic contents.

Limitation It is worth noting a few limitations of our method. The DFNet may not perform well when a viewing angle between neighboring views are larger (e.g., rotating more than 45°), which may decrease the amount of overlaps of dynamic contents. If the scene is highly cluttered by

Perceptual Sim. / Optical Flow ↘	Jumping	Skating	Truck	DynaFace	Umbrella	Balloon1	Balloon2	Teadybear	Avg.
MVS [43]+ASAPW [28]	0.21 / 7.0	0.17 / 9.3	0.10 / 4.0	0.30 / 19.0	0.19 / 7.5	0.23 / 16.0	0.17 / 6.7	1.80 / 4.9	0.19 / 9.3
RMVSNet [55]+ASAPW [28]	0.22 / 6.4	0.23 / 13.1	0.11 / 3.4	0.98 / 10.2	0.19 / 7.2	0.23 / 14.9	0.16 / 6.3	0.20 / 10.0	0.29 / 8.9
MonoDepth [23]+ASAPW [28]	0.23 / 9.1	0.18 / 11.8	0.10 / 5.1	0.32 / 20.9	0.20 / 9.8	0.25 / 17.3	0.23 / 11.4	0.17 / 7.8	0.20 / 11.7
Sparse2Dense [31]+ASAPW [28]	0.23 / 7.5	0.19 / 9.4	0.11 / 4.8	0.31 / 20.8	0.19 / 7.0	0.23 / 13.7	0.16 / 6.6	0.19 / 6.4	0.20 / 9.52
MVS [43]+B3W	0.24 / 7.0	0.20 / 9.2	0.12 / 3.5	0.27 / 7.5	0.19 / 5.7	0.23 / 14.4	0.17 / 5.4	0.13 / **1.5**	0.19 / 6.8
RMVSNet [55]+B3W	0.23 / 5.6	0.23 / 14.8	0.14 / 3.3	1.0 / 10.8	0.19 / 5.6	0.23 / 12.0	0.16 / 5.1	0.19 / 8.9	0.29 / 8.2
MonoDepth [23]+B3W	0.23 / 8.5	0.18 / 11.4	0.10 / 5.0	0.32 / 19.1	0.19 / 8.5	0.24 / 17.3	0.23 / 11.4	0.15 / 5.2	0.20 / 10.8
Sparse2Dense [31]+B3W	0.24 / 7.3	0.20 / 9.2	0.13 / 4.7	0.31 / 11.7	0.2 / 6.7	0.24 / 14.0	0.18 / 6.6	0.17 / 4.8	0.22 / 8.12
DFNet+ASAPW [28]	0.20 / 5.8	0.17 / 9.3	0.09 / 3.0	0.30 / 18.0	0.18 / 6.4	0.20 / 13.3	0.16 / 6.4	0.17 / 5.8	0.18 / 8.5
DFNet+B3W-DeepBlender	0.23 / 8.2	0.21 / 13.1	0.12 / 4.8	0.30 / 15.6	0.22 / 9.0	0.25 / 15.8	0.20 / 9.2	0.18 / 4.7	0.21 / 10.1
DFNet+B3W (**ours**)	**0.16 / 4.2**	**0.15 / 8.8**	**0.08 / 2.5**	**0.22 / 6.2**	**0.16 / 3.6**	**0.18 / 10.6**	**0.14 / 5.1**	**0.13** / 2.0	**0.15 / 5.3**

Table 2: Quantitative evaluation results on the dynamic scene novel view synthesis task. To measure the accuracy, we compute perceptual similarity and optical flow magnitude between the ground-truth and the synthesized image.

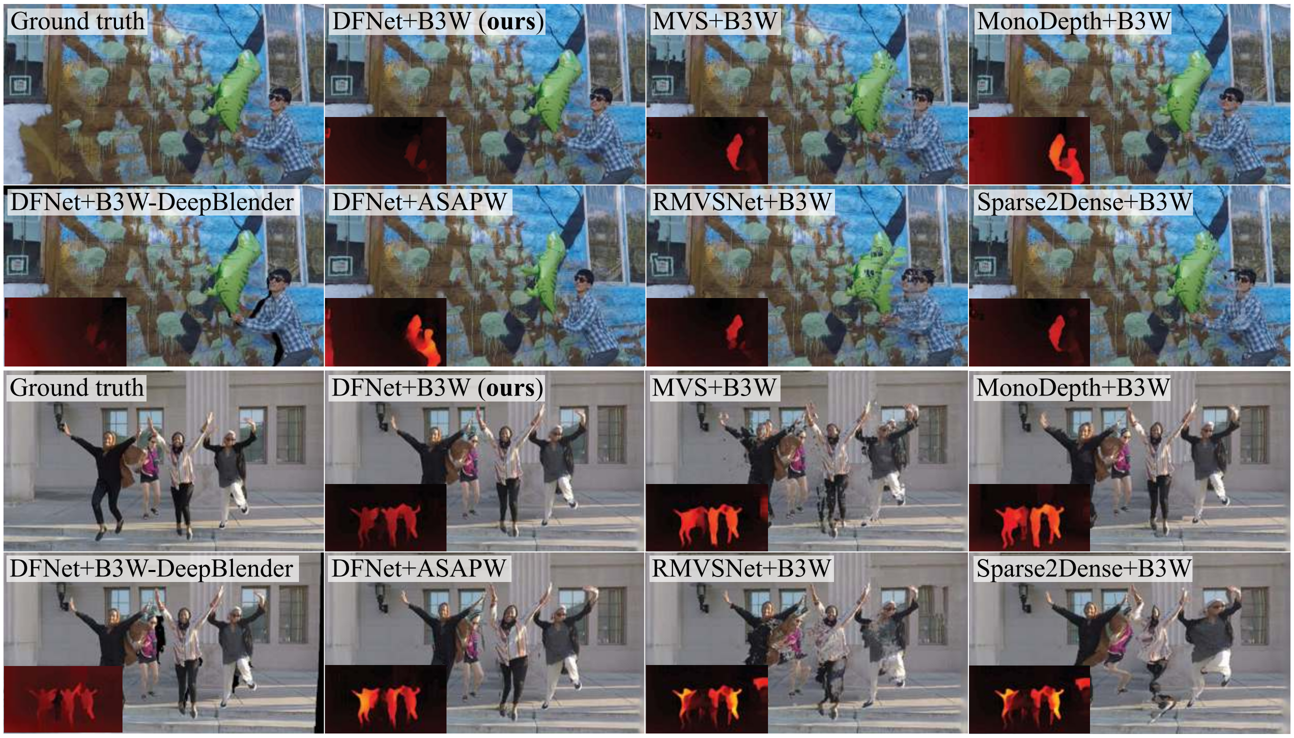

Figure 8: Qualitative comparison on the view synthesis task. The pixel error is shown in the inset image (maximum pixel error is set to 50 RGB distance).

many objects from both background and foreground (e.g., many people, thin poles, and trees), our pipeline could cause noisy warping results due to the significant depth discontinuities from the clutter. Our method will fail in the scenes where the camera calibration does not work, e.g., a scene largely occupied by dynamic contents [30]. Finally, our view synthesis with completely failed foreground mask produces significant artifacts such as afterimages and object fragmentation as shown in Figure 7.

6. Conclusion

In this paper, we study the problem of monocular view synthesis of a dynamic scene. The main challenge is to reconstruct dynamic contents to produce geometrically coherent view synthesis, which is an ill-posed problem in general. To address this challenge, we propose to learn a scale correction function that can upgrade the depth from single view (DSV), which allows matching to the depth of the multi-view solution (DMV) for static contents while producing locally consistent scene motion. Given the computed depth, we synthesize a novel view image using the DeepBlender network that is designed to combine foreground, background, and missing regions. Through the evaluations for depth estimation and novel view synthesis, we demonstrate that the proposed method can apply to the daily scenario captured from a monocular camera.

Acknowledgement This work was partly supported by the NSF under IIS 1846031 and CNS 1919965.

References

[1] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Nonrigid structure from motion in trajectory space. In *NIPS*, 2008.

[2] T. Alldieck, M. Magnor, B. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *CVPR*, 2019.

[3] Shai Avidan and Amnon Shashua. Trajectory triangulation: 3D reconstruction of moving points from a monocular image sequence. *TPAMI*, 2000.

[4] Jiamin Bai, Aseem Agarwala, Maneesh Agrawala, and Ravi Ramamoorthi. Automatic cinemagraph portraits. In *Proceedings of the Eurographics Symposium on Rendering*, 2013.

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

[6] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *NIPS*, 2019.

[7] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3D shape from image streams. In *CVPR*, 1999.

[8] Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *Conference on Computer Graphics and Interactive Techniques*, 1993.

[9] Inchang Choi, Orazio Gallo, Alejandro Troccoli, Min H Kim, and Jan Kautz. Extreme view synthesis. *arXiv preprint arXiv:1812.04777*, 2018.

[10] J. Fayad, Lourdes Agapito, and Alessio Del Bue. Piecewise quadratic reconstruction of non-rigid surface from monocular sequences. In *ECCV*, 2010.

[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*, June 2019.

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

[13] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, et al. Building rome on a cloudless day. In *ECCV*, 2010.

[14] Yasutaka Furukawa, Brian Curless, Steven M. Seitz, and Richard Szeliski. Manhattan-world stereo. In *CVPR*, 2010.

[15] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[16] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. *SIGGRAPH*, 2005.

[17] Dinghuang Ji, Enrique Dunn, and Jan-Michael Frahm. Spatio-temporally consistent correspondence for dense dynamic scene modeling. In *ECCV*, 2016.

[18] Hanqing Jiang, Haomin Liu, Ping Tan, Guofeng Zhang, and Hujun Bao. 3d reconstruction of dynamic scenes with multiple handheld cameras. In *ECCV*, 2012.

[19] Jeremy Yirmeyahu Kaminski and Mina Teicher. A general framework for trajectory triangulation. *JMIV*, 2004.

[20] Chen Kong and Simon Lucey. Deep interpretable non-rigid structure from motion, 2019.

[21] Johannes Kopf, Michael F. Cohen, and Richard Szeliski. First-person hyper-lapse videos. *SIGGRAPH*, 2014.

[22] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Monocular dense 3d reconstruction of a complex dynamic scene from two perspective frames. In *ICCV*, 2017.

[23] Katrin Lasinger, René Ranftl, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341*, 2019.

[24] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. *CVPR*, 2019.

[25] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018.

[26] Zicheng Liao, Neel Joshi, and Hugues Hoppe. Automated video looping with progressive dynamism. *SIGGRAPH*, 2013.

[27] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. Planercnn: 3d plane detection and reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4450–4459, 2019.

[28] Shuaicheng Liu, Lu Yuan, Ping Tan, and Jian Sun. Bundled camera paths for video stabilization. *SIGGRAPH*, 2013.

[29] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *SIGGRAPH*, 2018.

[30] Zhaoyang Lv, Kihwan Kim, Alejandro Troccoli, Deqing Sun, James M Rehg, and Jan Kautz. Learning rigidity in dynamic scenes with a moving camera for 3d motion field estimation. In *ECCV*, 2018.

[31] Fangchang Mal and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *ICRA*, 2018.

[32] S. Olsen and A. Bartoli. Using priors for improving generalization in non-rigid structure-from-motion. In *BMVC*, 2007.

[33] Hyun Soo Park and Yaser Sheikh. 3d reconstruction of a smooth articulated trajectory from a monocular image sequence. In *ICCV*, 2011.

[34] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders:feature learning by inpainting. In *CVPR*, 2016.

[35] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019.

[36] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016.

[37] Julien Philip and George Drettakis. Plane-based multi-view inpainting for image-based rendering in large scenes. In *SIG-*

GRAPH Symposium on Interactive 3D Graphics and Games, 2018.

[38] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019.

[39] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *SIGGRAPH*, 2004.

[40] C. Russell, J. Fayad, and L. Agapito. Energy based multiple model fitting for non-rigid structure from motion. In *CVPR*, 2011.

[41] Chris Russell, Rui Yu, and Lourdes Agapito. Video pop-up: Monocular 3d reconstruction of dynamic scenes. In *ECCV*, 2014.

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

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

[44] A. Shashua and L. Wolf. Homography tensors: On algebraic entities that represent three views of static or moving planar points. In *ECCV*, 2000.

[45] Sudipta N. Sinha, Drew Steedly, and Rick Szeliski. Piecewise planar stereo for image-based rendering. In *ICCV*, 2009.

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

[47] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018.

[48] Lorenzo Torresani and Christoph Bregler. Space-time tracking. In *ECCV*, 2002.

[49] Jack Valmadre and Simon Lucey. General trajectory prior for non-rigid reconstruction. In *CVPR*, 2012.

[50] Jack Valmadre and Simon Lucey. General trajectory prior for non-rigid reconstruction. In *CVPR*, 2012.

[51] Y. Wexler and A. Shashua. On the synthesis of dynamic scenes from reference views. In *CVPR*, 2000.

[52] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body and hands in the wild. In *CVPR*, 2019.

[53] Jing Xiao and Takeo Kanade. Non-rigid shape and motion recovery: Degenerate deformations. In *CVPR*, 2004.

[54] Jingyu Yan and Marc Pollefeys. A factorization-based approach to articulated motion recovery. In *CVPR*, 2005.

[55] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *CVPR*, 2019.

[56] Guofeng Zhang, Jiaya Jia, Tien-Tsin Wong, and Hujun Bao. Consistent depth maps recovery from a video sequence. *TPAMI, year=2009,*.

[57] Qi Zhang, Li Xu, and Jiaya Jia. 100+ times faster weighted median filter (wmf). In *CVPR*, 2014.

[58] 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.

[59] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018.

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