

Synthetic Defocus and Look-Ahead Autofocus for Casual Videography

XUANER ZHANG, University of California, Berkeley

KEVIN MATZEN, Facebook Research

VIVIEN NGUYEN, DILLON YAO, University of California, Berkeley

YOU ZHANG, Chapman University, Independent Filmmaker

REN NG, University of California, Berkeley



Fig. 1. We present a new approach to pursue cinema-like focus in casual videography, with shallow depth of field (DOF) and accurate focus that isolates the subject. We start with (A) a deep DOF video shot with a small lens aperture. We use a new combination of machine learning, physically-based rendering, and temporal filtering to synthesize (B) a shallow DOF, refocusable video. We also present a novel Look-Ahead Autofocus (LAAF) framework that uses computer vision to (C) analyze upcoming video frames for focus targets. Here, for example, we see face detection (white boxes) and localization of who is speaking/singing [Owens and Efros 2018] (heat map). The result is shallow DOF video (D), where LAAF tracks focus on the singer to start, and transitions focus to the child as the camera pans away from the musicians. The LAAF framework makes future-aware decisions to drive focus tracking and transitions at each frame. This presents a new framework to solve the fundamental realtime limitations of camera-based video autofocus systems.

In cinema, large camera lenses create beautiful shallow depth of field (DOF), but make focusing difficult and expensive. Accurate cinema focus usually relies on a script and a person to control focus in realtime. Casual videographers often crave cinematic focus, but fail to achieve it. We either sacrifice shallow DOF, as in smartphone videos; or we struggle to deliver accurate focus, as in videos from larger cameras. This paper is about a new approach in the pursuit of cinematic focus for casual videography. We present a system

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2019 Association for Computing Machinery.

0730-0301/2019/7-ART30 \$15.00

<https://doi.org/10.1145/3306346.3323015>

that synthetically renders refocusable video from a deep DOF video shot with a smartphone, and analyzes *future* video frames to deliver context-aware autofocus for the current frame. To create refocusable video, we extend recent machine learning methods designed for still photography, contributing a new dataset for machine training, a rendering model better suited to cinema focus, and a filtering solution for temporal coherence. To choose focus accurately for each frame, we demonstrate autofocus that looks at upcoming video frames and applies AI-assist modules such as motion, face, audio and saliency detection. We also show that autofocus benefits from machine learning and a large-scale video dataset with focus annotation, where we use our RVR-LAAF GUI to create this sizable dataset efficiently. We deliver, for example, a shallow DOF video where the autofocus transitions onto each person *before* she begins to speak. This is impossible for conventional camera autofocus because it would require seeing into the future.

CCS Concepts: • Computing methodologies → Computational photography; Image-based rendering.

Additional Key Words and Phrases: video-editing, autofocus, depth-of-field

ACM Reference Format:

Xuaner Zhang, Kevin Matzen, Vivien Nguyen, Dillon Yao, You Zhang, and Ren Ng. 2019. Synthetic Defocus and Look-Ahead Autofocus for Casual Videography. *ACM Trans. Graph.* 38, 4, Article 30 (July 2019), 16 pages. <https://doi.org/10.1145/3306346.3323015>

1 INTRODUCTION

Cinematic focus is characterized by the beautiful, shallow depth of field (DOF) of large lenses, which are prized for their ability to visually isolate movie stars, control the viewer’s gaze, blur out backgrounds and create gorgeous “bokeh balls” of defocused color. Focus that is tack sharp is essential, but shallow DOF makes it difficult and expensive to achieve. On a movie set, the primary camera assistant (“focus puller”) must operate the camera focus controls in realtime to track moving subjects and transition focus according to the screenplay. In movies with improvisational acting, such as *Coherence* (2014), the cinematographer must try to anticipate what the actors will do; of course significant focus error must be accepted. It is far more common to have a movie script, and for the focus puller to give actors markers on the ground to indicate where they should stand at specific points to facilitate highly accurate focusing.

These are the issues that make cinematic focus impossible for casual videographers, even though we would love to achieve the aesthetic. Instead, with smartphone videography we sacrifice cinematic DOF, because the small lenses cause essentially everything to be in focus at the same time. In contrast, with cameras that have larger sensors and lenses capable of cinema-like DOF, we inevitably sacrifice focus accuracy. The reason is that there is no movie script, so memorable moments and decisive actions occur unpredictably. Like the focus puller for the *Coherence* movie, the camera’s autofocus system would need a crystal ball to perfectly track each moving subject, or decide to transition focus to a new target in anticipation of its actions taking control of the narrative.

In this paper, we argue that a new direction is necessary if we are to truly deliver cinema-like focus for casual videography. An example unprecedented result we seek is a shallow DOF video of a group conversation where the focus transitions perfectly from person to person *before* each person begins talking. Another example is a video that faithfully tracks focus on a rapidly moving soccer player, and then presciently pulls focus onto another player before she heads the ball into the goal. The key to our new approach is to capturing refocusable video, and open the door to analyzing *future* video frames in order to deliver accurate tracking and anticipatory decisions about whether to transition focus to a decisive action by a new target. To enable these conventionally impossible capabilities, we contribute a framework composed of two modules.

(1) **Refocusable Video Rendering (RVR)** Rather than capturing regular videos with static focus, we produce synthetic “refocusable video” where the focus depth of each frame can be computationally changed after capture. Our approach is to synthetically render a shallow DOF video, from a deep DOF video that can be recorded with a smartphone. We build on recent methods in this vein, which

are limited to still photography and suffer from disturbing temporal inconsistency when applied frame by frame to videos. We extend synthetic shallow DOF to full video using a combination of machine learning, physically-based rendering and temporal filtering. For machine learning, we contribute a dataset of over 2,000 image pairs or triplets where the aperture and/or focus are varied, and use these data to train a convolutional neural network that improves prediction of RGBD video and recovery of HDR as input to Refocusable Video Rendering (RVR).

(2) Look-Ahead Autofocus (LAAF) for Casual Videography

We introduce the notion of Look-Ahead Autofocus that analyzes the seconds of video frames ahead of the current frame in order to decide whether to maintain or transition the focal depth. We demonstrate LAAF with examples of “AI-assistance” that include: motion and face detection to focus on upcoming human actions, audio localization to focus on who is about to speak, and a machine-learning-based focus detector that shows how a large-scale video dataset can be used to help autofocus of more generic videos. We build an interactive GUI incorporating with subject tracking and automatic focus transition so that the user only makes focus choices on a few keyframes to render a video with shallow DOF and annotated focus.

2 PRIOR ART AND RELATED WORK

2.1 Camera Autofocus Systems

Camera autofocus systems have generally been classified into two buckets: contrast-detection autofocus (CDAF) and phase-detection (PDAF). CDAF is slower, seeking focus by aiming to maximize image contrast as the lens focus is changed; it performs poorly for video because the “focus seeking” behavior is visible in the recorded video. Phase detection can be much faster, and is based on separately detecting and comparing light passing through different parts of the lens aperture. This was achieved in SLR cameras by reflecting light onto PDAF units that each comprised a microlens atop multiple pixels [Goldberg 1992].

To enable PDAF in mirrorless camera designs, sensor makers began embedding microscopic PDAF units sparsely into the pixel arrays themselves [Fontaine 2017], and advanced to the point that every imaging pixel became a PDAF unit to maximize light and autofocus-sensitive area [Kobayashi et al. 2016; Morimitsu et al. 2015]. This last design is now common in smartphones [Fontaine 2017; Levoy and Pritch 2017]. One might argue that such advances in physical autofocus systems are asymptotically approaching the fastest possible in many devices today. And yet, autofocus mistakes remain common and inevitable in casual videography, because the focus of each frame is “locked in” as it is being shot. A full solution is impossible because the autofocus algorithm would have to predict the future at every frame to correctly determine what to focus on, or transition focus to. This paper aims to lift this fundamental limitation by synthesizing shallow depth of field as a video postprocess, and using video autofocus algorithms that “look ahead” to make contextually meaningful predictions about what to focus on.

2.2 Light Field Imaging

Another way to capture refocusable images is a light field camera, but the cost of light field video systems remains very high. Rather

than capturing a 2D slice of the light field, as is the case with a conventional camera that samples a set of rays that converge at a single point, commercial light field cameras, e.g., Lytro ILLUM¹ capture the full 4D slice of the light field. Light field imaging not only captures the set of rays from different viewpoints [Levoy and Hanrahan 1996], but also enables physically-accurate synthetic aperture rendering and after-the-fact refocusing [Isaksen et al. 2000; Ng et al. 2005; Wilburn et al. 2005]. Beyond spatial resolution trade-offs, commercial light field video cameras currently have decreased video frame rate, approximately 3 FPS rather than the desired minimum of 30 FPS. Wang [2017] propose a hybrid system using one light field camera and one DSLR camera to produce 30 FPS light field video view interpolation.

2.3 Synthetic Defocus

Regarding synthesizing defocus in images, previous work has achieved breakthroughs on still photography and led to common “portrait” modes in current smart phone photography that synthetically blur the background in portrait photography. To the best of our knowledge, ours is the first to push such synthetic defocus from still photography to full videography. The first step towards RVR is to accomplish single image synthetic defocus, we give a high-level review on synthetic defocus for still images.

Stereo can be used to derive necessary depth [Joshi and Zitnick 2014; Yu et al. 2011]. The dual pixel sensors described earlier can be used to provide stereo views of the scene through right and left halves of the lens aperture. This has been used to estimate depth for synthesizing defocus blur for smartphone computational photography [Levoy and Pritch 2017; Wadhwa et al. 2018].

Data-driven machine learning approaches have also proven valuable in synthetic defocus tasks using single images, a lot of which are for driving scenarios using specialized datasets [Garg et al. 2016; Godard et al. 2017; Kuznetsov et al. 2017]. MegaDepth [Li and Snavely 2018] is a concurrent work that targets more generic image contents. Most related to our own work is the method of Srinivasan *et al.* [2018], who use weak supervision by predicting a depth map for an input photo, passing it through a differentiable forward rendering model, and then applying a reconstruction loss to the output shallow depth-of-field photo. However, they employ a simplified forward model and avoid having saturated pixels (e.g. salient bokeh regions) in their dataset. Park *et al.* [2017] combines hand-crafted features with deep features to render refocusable images. However, they focus on noticeable defocus that is generated by medium-large aperture sizes. Therefore their method degrades on images taken by ~f/16 and smaller aperture sizes, while this paper focuses on deep DOF input videos generated by f/20 or smaller.

A number of works also use defocus as a cue to predict depth of the scene [Mather 1996]. Nayar and Nakagawa [1994] propose a focus operator which compares texture variability between images to determine relative level of focus. Correspondences from light-field imaging can be used to reduce ambiguity in depth-from-defocus [Tao et al. 2013]. Suwajanakorn *et al.* [2015] recently explored depth-from-defocus for smart-phone imagery and Tang *et*

al. [2017] generalize depth-from-defocus for unconstrained smart-phone imagery in the wild using two perceptibly similar images.

2.4 Video Analysis

Understanding video contents, such as knowing when and where activity happens or which regions are visually salient, is key to our video autofocus algorithm. Recent advances in video understanding tasks such as activity classification [Karpathy et al. 2014; Soomro et al. 2012], activity recognition and detection [Bilen et al. 2016; Feichtenhofer et al. 2016; Sung et al. 2012], benefit LAAF in localizing action in videos. Video saliency detects salient subjects under a more generic context, often making use of eye tracking signals [Sun et al. 2018; Wang et al. 2018a]. We find saliency detection effective in proposing a coarse focus region for LAAF, and it can be combined with other detectors for finer-grained localization. A topic similar to video attention along the temporal axis is video summarization [Mahasseni et al. 2017; Zhang et al. 2016], extracting keyframes or subshot as visual summaries for long videos. More recently, audio-visual signals have been combined jointly to learn semantically meaningful video representations, one application we use to detect and locate people who are speaking is audio source localization [Ephrat et al. 2018; Owens and Efros 2018].

3 SYSTEM OVERVIEW: RVR-LAAF

Our system (See Figure 2) aims to render cinematic autofocus for casual videos. It consists of two components: Refocusable Video Rendering (RVR) (Section 4) – rendering videos with shallow DOF focused at any depth at any time, and video Look-ahead Autofocus (LAAF) (Section 5) – choosing when and where to focus to make the autofocus choices contextually meaningful and visually appealing.

RVR is built upon a refocusable single frame renderer and a temporal module. We summarize our contribution in rendering refocusable video in Figure 3. We find it key to render RVR with temporal coherence, estimation of HDR detail, and a physically-based forward model of lens defocus. We achieve temporal stability by applying an occlusion-aware temporal filtering that is based on optical flow and robust to outliers (see Section 4.3 and Figure 3A). For photo-realistic rendering, we train a neural network to jointly estimate, from a single image, the defocus size and unclipped intensity value for each pixel. We find HDR recovery enables rendering of realistic bokeh (Figure 3B) and a correct forward model enables correct occlusion effects (Figure 3D). To train the network, we collect a large-scale aperture dataset that contains image pairs and triplets. We find our collected triplet dataset to improve estimation around large disparity regions (Figure 3C). We call our trained network a RGBD-HDR estimator (Section 4.1). RVR takes a deep DOF video that can be captured by a smartphone, and generates a shallow DOF video that can be focused on any depth at any frame.

RVR delivers video that can be focused at any depth, but the question remains: what is the correct depth to focus on at every frame? For example, retaining optical focus and simply synthesizing shallow DOF (see Figure 4C row 2, and the video at 03:18) results in obvious focus errors and often lacks contextual meaning (e.g. focusing on one person while another person, blurred out, is speaking).

¹https://www.dpreview.com/products/lytro/compacts/lytro_illum

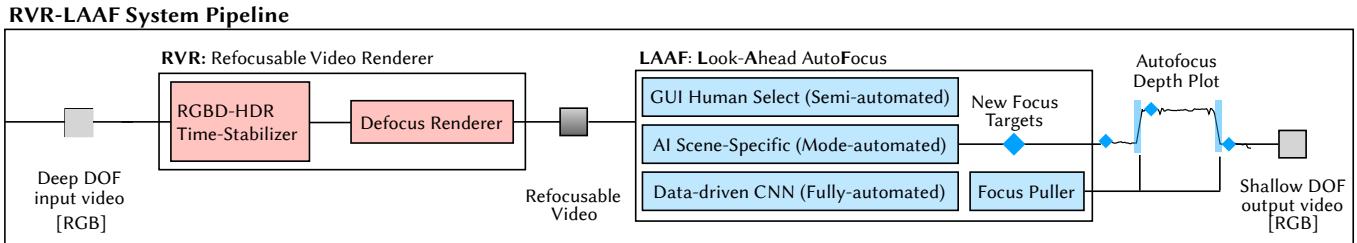


Fig. 2. Overall system pipeline to compute shallow DOF video from a deep DOF video input. The Refocusable Video Renderer (RVR) contains a time-Stabilized RGBD-HDR (Section 4) that computes temporal-stabilized depth and recovers HDR. The Look-Ahead Autofocus (LAAF) (Section 5) pipeline consists of three approaches with different levels of automation, detecting New Focus Targets and generating autofocus depth followed by a focus puller that smooths focus transition. Output is a shallow DOF video with contextually-meaningful focus.

Our solution is called the LAAF framework and comprises three complementary techniques for attacking this problem of “when and where to focus”. First (Section 5.1), LAAF contains a carefully designed user interface (RVR-LAAF GUI) that enables a user to specify only a small number of semantically meaningful “new focus targets” in a video clip – the system then tracks these subjects to maintain focus on them, and adds focus-pull transitions automatically. Second (Section 5.2), LAAF provides AI-based autofocus modules that only requires the user to choose the type of scene – then, the system fully automates the task by choosing “new focus targets” intelligently. We demonstrate examples for scenes that contain conversations or actions, where the system automatically pre-focuses on each person before they speak or on actions before they occur. Third (Section 5.3), (Section 5.3), we demonstrate a first attempt at fully automating video autofocus using a machine learning approach – we contribute ground-truth focus annotations on a large-scale video dataset, using our RVR-LAAF GUI to create this sizable dataset efficiently.

4 REFOCUSABLE VIDEO RENDERING (RVR)

In this paper we introduce the first synthetic defocus renderer for videos. We take Monocular Depth Estimation (MDE) [Srinivasan et al. 2018] as our launching point, and comprehensively re-work the method to make it suitable for synthesizing defocus for casual videography. This section presents the implementation details. The four critical changes are:

- (1) Adding HDR recovery estimation, which we show is critical to achieving plausible bokeh balls that are a visual hallmark of shallow DOF videos (Figure 3B, Section 4.1)
- (2) Building a superior training dataset with far greater scene diversity. This dataset contains novel “triplets” of images (Figure 5) that we find is important to improve depth estimation in foreground and background regions (Figure 3C Section 4.1).
- (3) Correcting the forward model to make it handle occlusions correctly and so that background defocus does not incorrectly “bleed” around foreground objects (Figure 3D, Section 4.1)
- (4) Adding temporal coherence to ameliorate flicker (Figure 3A, Section 4.3)

4.1 RGBD-HDR Estimator

We offer an indirectly-supervised approach to estimate disparity and HDR with only aperture supervision, by training a neural network

that jointly predicts a disparity map \mathbb{D} and recover high dynamic range (HDR) \mathbb{E} from a single image that has a deep DOF and standard dynamic range (SDR). \mathbb{D} and \mathbb{E} will be formally defined later. Unlike the MDE dataset that avoids saturation problems by avoiding saturated pixels in the input photo, our extended MDE dataset has a diverse set of scenes that cover a wide dynamic range and contain saturated pixels that are common in casual videography. We find that both disparity and HDR can be estimated purely from our extended MDE dataset, by imposing supervision on the reconstructed shallow DOF images. Alternative to our joint prediction is to estimate depth and HDR separately in sequence, for example, an HDR recovery network followed by a depth prediction network. Recent works have addressed performance on monocular depth estimation [Li and Snavely 2018] and monocular HDR recovery [Eilertsen et al. 2017a]. Our sub-system on monocular shallow DOF rendering would be approximately equivalent to a composition of these state-of-the-art works. One of our advantages is that we do not require ground truth supervision on either depth or HDR, while the aforementioned methods use direct supervision and thus require challenging data capture and annotation to account for model generalization to different or more generic scene content.

We formulate our joint disparity and HDR estimation network as following. Given training dataset \mathcal{A} with pairs of small aperture input I^S and large aperture ground truth output I^L , the major loss we apply to optimize the network parameters is the rendering loss:

$$L_{\text{rend}} = \|I^L - \mathcal{F}(I^S, \mathbb{D}, \mathbb{E})\|_1 \quad (1)$$

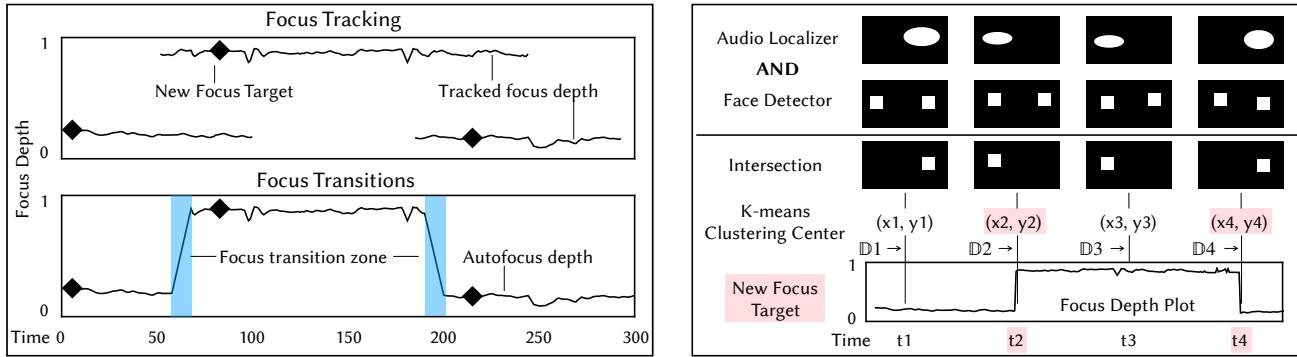
We use shallow DOF rendering as the objective, thereby bypassing the need to have direct supervision using disparity or HDR ground truth, which can be extremely challenging to collect and annotate.

The forward model \mathcal{F} is based on an ideal thin lens model [Potmesil and Chakravarty 1982]. It takes a deep DOF image with predicted \mathbb{D} and \mathbb{E} to render a synthetic shallow DOF image. We use a disk kernel K to approximate the point spread function of a defocused point through the lens. To handle occlusion, we blend layers of different disparity levels in order from back to front to prevent background blur from incorrectly bleeding around the silhouette of foreground objects, as shown in Figure 3D. Previous methods such as [Srinivasan et al. 2018] simply sum up all disparity levels.

We define disparity \mathbb{D} , the inverse depth, in its stereo sense being created by the differing vantage points from left and right edges of

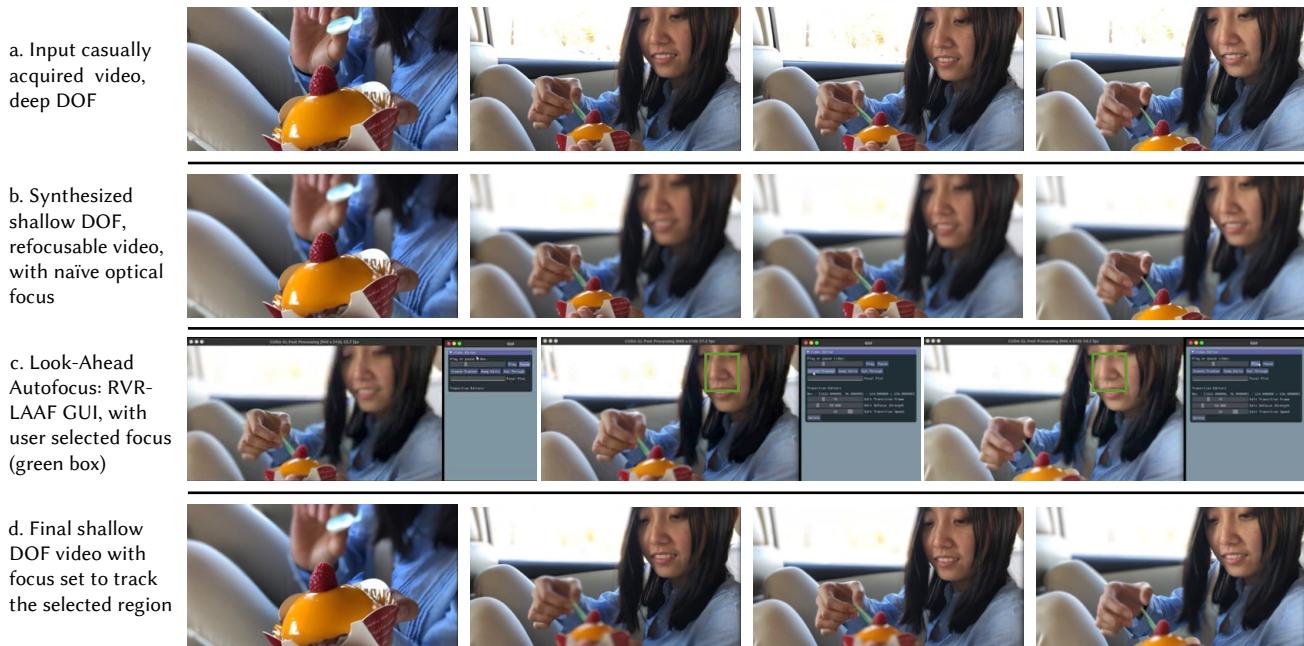


Fig. 3. Summarized contribution of our Refocusable Video Rendering subsystem (Section 4).



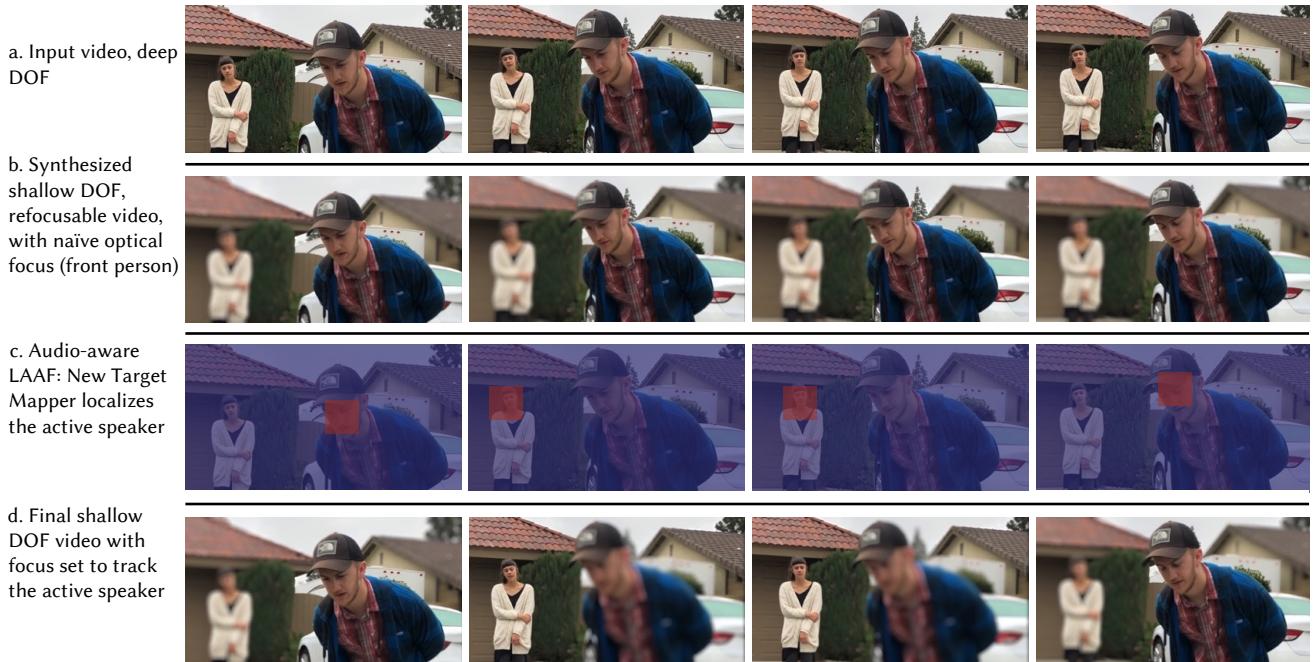
(4A) Illustration of synthetic focus pulling. Input to the module is a set of $\{(x, y, t)\}$ triplets denoting new focus targets and times. Three such focus targets are shown as diamonds on the timelines above. For each of these targets, we perform focus tracking (top graph) by computing (x, y) tracking across time, and look up the focus depth from the estimated depth map. Next, we execute fast focus pull transitions (bottom graph) from one target to the next, with focus arriving at each target slightly before t to allow the viewer to visually settle before action begins.

(4B) Illustration of computing New Focus Targets from scene-specific LAAF. For a given scene (eg. conversation), a set of detectors generate focus probability maps and their intersection is computed. K-means clustering finds the center of the majority cluster and generates its center (x, y) , which reads from predicted disparity map to produce a focus depth plot. We apply 1-D bilateral filter to the plot and detect focus depth discontinuity to be the New Focus Targets, denoted as (x, y, t) marked in pink for this example.



(4C) GUI-based video autofocus. (Row 1) Input video shows the camera shifts from the mango tart to the person. (Row 2) naïve synthetic SDOF video would not be able to resolve focus in time, blurring out the person's face. (Row 3) Our RVR-LAAF GUI is incorporated with vision-based tracking. (High-res GUI demo can be found in the accompanying video) User only needs to annotate New Focus Target (Row 3 Middle) when they identify a preferred focus subject. The GUI tracks the subject (Row 3 Right) until the next New Focus Target is selected. (Row 4) The resulting video shows autofocus that transitions correctly to the face without lagging, as the user knows camera is shifting to a more preferred focus subject. The video result can be found in the accompanying video.

Fig. 4. Summarized contribution of our Look-Ahead AutoFocus subsystem (Section 5).



(4D) Audio-aware video autofocus. (Row 1) Input video shows two people discussing a bug on the ground. (Row 2) naïve synthetic SDOF video would focus on the foreground person exclusively, blurring out the speaker in the background. (Row 1 sub) Audio localization mapping, a recent advance in computer vision, identifies the active speaker at each point in the video. (Row 4) The resulting video shows autofocus that transitions smoothly from speaker to speaker, before they begin speaking. The video result can be found in the accompanying video.



(4E) Action-aware video auto-focus. (Row 1) Input video shows children playing in the background and plant in foreground, optically focused on foreground. (Row 2) Naïve synthetic SDOF video would blur out the action in the background, because the input video is optically focused on the foreground. (Row 3) Intersection of visual saliency mapping and motion detection mapping isolates action of children in the background. (Row 4) The resulting video is autofocused on the background.

Fig. 4 (Cont.). Summarized contribution of our Look-Ahead AutoFocus subsystem (Section 5).

the lens aperture. This amount is proportional to the defocus blur size in pixel space by a normalized scalar [Held et al. 2010]. Assume the disparity map \mathbb{D} ranges from d_{\min} to d_{\max} and denote disk kernel radius as r ($r = 0$ at focal plane). We discretize \mathbb{D} into $|d_{\max} - d_{\min}|$ levels using a soft mask M . We define M as the matte for content at the corresponding disparity level. M allows the forward model \mathcal{F} to be differentiable and also stabilizes training. Formally, M at disparity d is defined as:

$$M(\mathbb{D}, d) = \exp(-\lambda(\mathbb{D} - d)^2) \quad (2)$$

λ is empirically set to 2.8 to model a continuous and rapid falloff across neighboring disparity levels. Because we predict signed disparity and define focal plane to have zero disparity, $r = d(d \geq 0)$ when d is behind the focal plane, otherwise, $r = -d(d < 0)$. We use I^l to denote the rendered shallow DOF image. Using our back-to-front forward rendering model, the shallow DOF image at disparity level d , denoted as I_d^l , is computed by blending it with its previous disparity level I_{d-1}^l using the aforementioned content mask M :

$$I_d^l = I_{d-1}^l \cdot (1 - M(\mathbb{D}, d)) + (I^S \cdot M(\mathbb{D}, d)) \otimes K(r) \quad (3)$$

One of the beautiful and characteristic visual signatures of shallow DOF video are so-called "bokeh balls," which are bright disks optically created by defocusing of small, bright lights (see the background of the middle image in Figure 5 and 02:03 in the video). A numerical challenge for synthesizing such bokeh balls is that one needs HDR images to capture the very high intensity of the small lights [Debevec and Malik 2008]. Without HDR, the represented intensity of the small bright lights is limited by the 8-bit fixed point precision of our input videos – when synthetically defocusing these lights, the incorrectly low intensity value spreads out over many pixels and becomes invisible rather than a bright ball of light (see middle image in Figure 3B). Therefore, we find HDR recovery key to render visually salient bokeh that appears at saturated regions, which has not been considered in prior synthetic defocus rendering models [Kraus and Strengert 2007; Wadhwa et al. 2018; Yang et al. 2016]. We undo gamma correction on input images to work in linear space. We predict \mathbb{E} in log scale to recover a high dynamic range. Pixels that are saturated in the deep DOF image are often not saturated in its shallow DOF pair, because their energy is spread over many pixels. This provides indirect signal for the network to learn HDR recovery. We replace I^S in Equation 3 with its HDR version $I^{S'}$, computed as:

$$I^{S'} = I^S \cdot e^{k \cdot \mathbb{E}} \quad (4)$$

where k affects the maximum recovered saturation value and is empirically set to 50.

Data Collection. To train our RGBD-HDR Estimator, we build upon the Flower dataset [Srinivasan et al. 2018] and contribute the first large-scale aperture dataset that covers diverse object categories. The dataset contains 1.2K image pairs and 0.8K image triplets taken with different aperture sizes (f/2, f/8 and f/22) and focus depth. Each image pair or triplet is taken in a scripted continuous shot using Magic Lantern² firmware add-on for Canon EOS DSLR cameras. This minimizes misalignment among pairs/triplets during capturing time.

²<https://magiclantern.fm/>

Pixel-wise alignment is further imposed via correlation coefficient minimization [Evangelidis and Psarakis 2008].



Fig. 5. Example image pairs and triplets in our dataset. (Row 1) A focal triplet example, defocus map predicted from the input image is used to reconstruct the image with the same focus depth but taken with a large aperture (middle image) shifted to reconstruct the image taken with a large aperture at a different depth plane (right image). (Row 2) An example of aperture triplet, defocus map predicted from the input image is used to reconstruct the large aperture image, and scaled to reconstruct the medium aperture image. (Row 3) An example of image pairs, note that the focus is on the middle plane for this example.

Image Pair and Triplet Supervision. The network takes in a deep DOF image and predicts a disparity map and a high dynamic range, which are used to render a shallow DOF image, and has ground truth to compare against. This rendering loss is back-propagated to update network parameters until convergence. We notice that the precision of large defocus values in \mathbb{D} is less accurate as the gradient of the reconstruction loss decreases inversely proportional to the size of the defocus disk kernel ($\frac{\partial L_{\text{rend}}}{\partial r} \propto \frac{1}{r^2}$). This produces visual artifacts when refocusing the video to planes that are originally at large disparity. To mitigate imbalanced loss gradient back-propagated through different disparity planes, we apply a triplet consistency checking during training. Our dataset contains two types of image triplets: aperture and focal triplets. *Aperture triplets* are taken with f/2, f/8 and f/22. The estimated disparity should be able to *scale* to render both median DOF and shallow DOF images. This constraint also helps stabilize training. In contrast, *focal triplets* include a deep DOF image and two shallow DOF images focused at different depths. The estimated disparity map should be able to *shift* to render both shallow DOF images at different depths. From our thin lens model assumption, the non-linearity between change of focal plane and change of defocus blur size only depends on the object disparity and lens movement in sensor coordinate, which is relatively small to be negligible. As long as the object is not too close to the camera, we can assume the shift of focal plane to be linear to the change of defocus size. For the prime lens, Canon EF 50mm f/1.8, that we use for data capture, scene depths difference from infinity to 0.5m generate a deviation of ~5% from the assumed linear model. Training with both types of triplet data helps to improve the precision of large

disparity region estimation. Figure 3C shows a visual comparison on the estimated disparity map and back-focus shallow DOF rendering between training with and without triplet consistency data.

Loss Functions. We train RGBD-HDR with rendering objectives that penalize difference between rendered results and ground truth shallow DOF targets. We supervise the network with per-pixel L^1 loss, denoted as L_{pix} , as well as low-level and high-level image features denoted as L_{feat} , by feeding the network output and target through a pre-trained VGG-19 network Φ . We compute the L^1 difference between $\Phi(\mathcal{F}(I^S, \mathbb{D}, \mathbb{E}))$ and $\Phi(I^L)$ in selected feature layers.

$$\begin{aligned} L_{\text{rend}}(I^S, I^L) &= L_{\text{feat}}(I^S, I^L) + L_{\text{pix}}(I^S, I^L) \\ &= \sum_i \lambda_i \|\Phi_i(I^L) - \Phi_i(\mathcal{F}(I^S, \mathbb{D}, \mathbb{E}))\|_1 + \\ &\quad \|I^L - \mathcal{F}(I^S, \mathbb{D}, \mathbb{E})\|_1, \end{aligned} \quad (5)$$

where Φ_i indicates the layer i in the VGG-19 network. The weights $\{\lambda_i\}$ are used to balance different terms in the loss function. We select the layers conv1_2, conv2_2, conv3_2, conv4_2, and conv5_2 in the VGG-19 network. These features are demonstrated to be effective for image enhancement, style transfer and many other image processing tasks [Chen and Koltun 2017; Johnson et al. 2016; Zhang et al. 2018].

Image triplets are applied in an adjusted rendering objective that penalizes difference from shallow DOF image I^L after adjusting aperture (scale) and focal plane (shift), with the adjustments linearly approximated by an affine model. Similar to L_{rend} , $L_{\text{adjust_rend}}$ computes the L^1 difference between $\Phi(\mathcal{F}(I^S, \alpha\mathbb{D} + \beta, \mathbb{E}))$ and $\Phi(I^L)$ in the same selected feature layers. We find this image triplet training effectively improves RGBD-HDR prediction at far planes, as shown in Figure 3C.

We additionally incorporate an edge-aware smoothness penalty L_{smooth} on the predicted disparity map by minimizing the L^1 norm of its gradients, weighted less on the input image edges. This ensures the predicted disparity map to be locally smooth and is formulated as:

$$L_{\text{smooth}}(\mathbb{D}) = \|\partial_x \mathbb{D}\|_1 \cdot e^{-|\partial_x I^S|} + \|\partial_y \mathbb{D}\|_1 \cdot e^{-|\partial_y I^S|}$$

Overall, we train our network by minimizing a loss function that is a weighted sum of L_{rend} , $L_{\text{adjust_rend}}$, and L_{smooth} .

$$\begin{aligned} L_{\text{total}} = \sum_{(I^S, I^L, I^{L'}) \in \mathcal{A}} & L_{\text{rend}}(I^S, I^L) + L_{\text{adjust_rend}}(I^S, I^{L'}) \\ & + w_1 L_{\text{smooth}}(\mathbb{D}) \end{aligned} \quad (6)$$

where w_1 is the weight for the smoothness regularization, and is set to 10 across all experiments. When a triplet is not available for a particular example, i.e., only a pair is available, we omit $L_{\text{adjust_rec}}(I^S, I^{L'})$ and double the weight of $L_{\text{rend}}(I^S, I^L)$.

Training and Implementation. We use a U-net network architecture [Ronneberger et al. 2015] that contains an encoder-decoder structure with skip connections. All layers are followed by a leaky *ReLU* activation, except for the last prediction layer that produces $3 + N$ channels. Three of these channels are used for HDR recovery. The other N channels are used as a bilateral-space representation over luma, which are sliced with a bilateral slicing operator (where

N is defined by bandwidth parameters of the bilateral slicing operator) into pixel-space to produce a 1-channel signed disparity map. In practice, we find that predicting a bilateral-space representation improves fidelity of the disparity map over predicting directly in pixel-space, particularly around edges. This is consistent with the findings of Gharbi et al. [2017] and Barron et al. [2015]. Positive disparity refers to planes behind the focal plane while negative disparity represents planes in front of the focal plane.

We train the network with batch size 1 on an NVIDIA Titan X GPU and weights are updated using the Adam optimizer [Kingma and Ba 2015] with a fixed learning rate of 10^{-4} . A full network architecture will be made available in a code release. The network converges after 150K iterations. Our network is fully convolutional and can run at arbitrary image sizes. During training, we resize the images to random resolutions between 512p and 1024p.

4.2 Video Temporal Consistency.

Visually, we find that the most important change when going from still images to video is to enforce temporal consistency. Independently rendering each frame with shallow DOF causes visually disturbing flickering, especially around prominent bokeh. A comparison can be found in Figure 3A and 01:44 in the video. To impose temporal coherency, we apply a weighted temporal moving average that is occlusion-aware and robust to outliers to \mathbb{D} and \mathbb{E} . For each target frame I_i , we compute $w_i, w_{i-1}, \dots, w_{i-M} \in W$, where M is the number of neighboring frames. We compute optical flow using a pre-trained deep neural network Flownet 2.0 [Ilg et al. 2017] for consecutive pairs of frames, and align I_{i-1}, \dots, I_{i-M} to I_i using concatenated flows. w_i is computed as a weighted combination of an occlusion weight w_i^{occl} and an outlier weight w_i^{med} .

Occlusion Weight. Occluded pixels should be weighted little. We adopt the tactic of forward-backward consistency checking [Chen and Koltun 2016; Ogale et al. 2005], computing both forward $f^{i \rightarrow i+1}$ and backward optical flow $f^{i+1 \rightarrow i}$ for frame pair I_i and I_{i+1} . Consider point p in f_i that shifts to $p + f^{i \rightarrow i+1}(p)$. We check if we can find a point q in f_{i+1} such that:

$$\|(p - f^{i+1 \rightarrow i}(q))\|_2 + \|(q - p + f^{i \rightarrow i+1}(p))\|_2 \leq \delta \quad (7)$$

where δ is a small distance threshold. If there exists such a q in f_{i+1} , p is considered as a consistent pixel in f_i . For each frame I_i we compute such an occlusion mask and call it w_i^{occl} .

Outlier Rejection. To account for optical flow inaccuracies, we classify a warped pixel as an outlier if it has very different values within its temporal moving window. We assume that outliers are sparse and thus a majority vote approach such as median filtering would be effective. For any point p in frame I_i , its outlier weight $w(p)$ is set to $e^{-\text{median}(|p_i + f^{j \rightarrow i}(p) - p_j|_{j=i-M}^i)}$. We compute an outlier weight for each frame and call it w_i^{med} . Overall, the filtered prediction \mathbb{D}_i (or \mathbb{E}_i) of target frame I_i is computed as:

$$\mathbb{D}_i = \sum_{j=i-M, \dots, i-1} W_j \cdot (\mathbb{D}_i + f^{j \rightarrow i}) + W_i \cdot \mathbb{D}_i \quad (8)$$

where $\sum_{i-M}^i W_j = 1$. We set M to 6 for all experiments.

4.3 Parallelization in Scanner.

We used a set of 100 test videos in developing and validating our prototype, so we use modern infrastructure to process them with RVR. We choose to use Scanner [Poms et al. 2018], which gives us the option to process on different hardware and parallelize on the cloud. With full parallelism we could in principle process all 100 videos in under a minute. In practice we use a local 4-core machine with a single Titan X GPU. On average, we process one 2 megapixel video frame in ten seconds, including RGBD-HDR inference, bi-directional optical flow computation, and temporal filtering. The current bottleneck is the flow-based temporal filtering and the fact that we do not optimize the total processing time to pursue full advantage of Scanner’s distributed processing of jobs.

5 LOOK-AHEAD AUTOFOCUS (LAAF)

Being able to synthesize refocusable videos is not a complete solution to generating a meaningful shallow DOF video – deciding on when and where to focus in the video is challenging and essential. In the conversation example shown in Figure 4D, the focus should shift between the active speaker and lock onto the person about to speak next, *before* she speaks. On a movie set, exact focus is achieved by a movie script that exhaustively defines what should be in focus at every moment in the film, and a dedicated focus puller (the 1st Assistant Camera) who measures and marks the exact focus position. In this section, we demonstrate how LAAF uses recent computer vision advances in video understanding – analyzing semantics of current and future frames – to enable video autofocus that automates portions of focusing process in cinematography. Our prototypes demonstrate, we believe for the first time, that it is actually tractable to deliver. We demonstrate three approaches towards semi-automated and fully-automated video autofocus, using:

- (1) an interactive GUI with vision-based tracking and simple human focus selection (Section 5.1)
- (2) scene-specific (e.g. conversation, action, etc.) AI-assist modules (Section 5.2)
- (3) a data-driven CNN network trained from a large-scale video dataset with focus annotation, labeled using our GUI (Section 5.3)

Output from the above three approaches is a set of $\{(x, y, \hat{t})_i\}$ triplets denoting New Focus Targets – focus regions and times. Three such New Focus Targets are shown as diamonds on the timelines in Figure 4A. For each New Focus Target, we perform focus tracking by computing (x, y) tracking across time, and look up the focus depth or subject from the estimated depth map. Next, we execute a digital focus pull from one target to the next, with focus arriving at each target slightly before the target time, to allow the viewer to visually settle before action begins. We set a default duration of 10 frames (~ 0.67 sec) for a focus pull and linearly interpolate focal planes in between.

For real systems, LAAF could happen during video capture as well. To do this, we would need to buffer a few seconds of video frames, which would be the temporal window we set to look ahead, and then pipeline the LAAF processing with video recording.

5.1 GUI-based Video Semi-Autofocus

– Human Selection of New Focus Targets

We build an interactive RVR-LAAF GUI incorporating a vision-based tracker (e.g. KCF tracker [Henriques et al. 2015]) such that the user only needs to specify a New Focus Target, instead of selecting a focus subject for each frame, which is extremely inefficient and impractical. The tracker follows the selected focus region until the user pauses the video to select the next New Focus Target (see 03:08 in the video).

One interesting point is that RVR-LAAF GUI provides benefit even for simple scenes that seem amenable to conventional autofocus. One might think that in these situations simply synthesizing shallow DOF from the recorded video would suffice. However, the issue is that synthetic defocus will amplify any misfocus error. For example, a subtle defocus from a person’s eyes to back of the head would be almost unnoticeable in a deep DOF input image, but after synthetic defocus the person’s features would become unacceptably blurred out – in this situation LAAF can keep the focus locked maintain sharp focus on the person’s features throughout the video. So even in these situations, RVR-LAAF GUI allows us to increase the focus accuracy of the output video by adjusting the synthetic focal plane onto the person of interest and track the corrected subject. In other words, autofocus achieved by LAAF is essential to delivering even simple synthetic defocus video accurately.

5.2 Scene-Specific Video Autofocus

– AI-Based Selection of New Focus Targets

Fully-automated video autofocus without human interaction requires visual and semantic understanding of the video context. We show how LAAF incorporates recent advances in video understanding to automate, to some extent, human choices of focus selections. Faces, actions, audio sources and other salient (visually distinctive) objects, are some common subjects to set in focus in casual videography. We exploit a set of context-aware detectors \mathcal{H} (e.g. face, action detectors) to automate the generation of New Focus Targets.

As illustrated in Figure 4B, we compute the intersection of a selected set of detection to identify a scene-dependent and contextually-meaningful focus region. We then apply K-means clustering (K empirically set to 4) to determine the majority clustering centroid position (x, y) and read in the focus depth from the predicted disparity map \mathbb{D} . Frames with empty intersection use its previous disparity level. Next, we apply a bilateral filter followed by an edge detection to identify focus depth discontinuity, which marks the New Focus Target, (x, y, t) . Note that we use (x, y) instead of $\mathbb{D}(x, y)$ to denote New Focus Target because we later use (x, y) to track the subject to synthesize focus puller as shown in Figure 4A.

For example, we demonstrate scene-specific LAAF on two types of videos, one on action videos that use action and saliency detection ($\mathcal{H} = \{H_{act}, H_{sal}\}$) to detect salient regions that also involve action (Section 5.2); one on conversation videos that use audio localization and face detector ($\mathcal{H} = \{H_{aud}, H_{face}\}$) to detect and focus on the person who is speaking (Section 5.2). These two types of videos are commonly seen in both casual videography and professional filmmaking. According to a video essay that attempts to compile a list of significant examples of focus racking in film history(years

1963–2016)³, half of the focus racks are triggered by human action or audio source change.

Action-aware LAAF. We collect a set of videos that involve unexpected actions, which trigger focus depth change towards the action subject. An example scenario is a background subject of interest entering the frame unexpectedly while focus is at the foreground, and the result we seek is to shift focus at, or a few frames before, the subject entering the view. We use an action localizer (H_{act}) and a salient object detector (H_{sal}) to generate probable focus regions that are both salient and involve actions. H_{act} is based on optical flow and deepmatch [Revaud et al. 2016] for action localization and tracking, and is used as video pre-processing step for several computer vision tasks such as unsupervised video feature learning [Pathak et al. 2017]. H_{sal} is based on a still image saliency detection [Hou et al. 2017] work, which trains a deep network to compute a saliency heat map that identifies visually distinctive objects and regions in an image. As illustrated in Figure 4E and 05:22 in the video, LAAF analyzes future frames (Row 1) and detects the child’s action that is about to happen. Instead of always keeping focus on the foreground bush, LAAF is able to shift focus to the background before the child slides down the hill (Row 4).

Audio-aware LAAF. We demonstrate LAAF applied to another video collection of conversational scenes. The goal is to place focus right before the person who is about to speak. We use an audio localizer (H_{aud}) and face detector (H_{face}) to compute probable focus regions of the person who is speaking. H_{aud} employs a recent breakthrough on audio localization [Owens and Efros 2018], where the authors train a deep network to learn multisensory representation using the fact that visual and audio signals often align temporally. H_{face} is a machine-learning-based face detector from dlib⁴. As illustrated in Figure 4D and 04:27 in the video, without LAAF the focus is always on the front person, blurring out the person in the back even when she starts talking (Row 2). LAAF analyzes future frames to understand who’s speaking, and is able to correctly shift focus a few frames before the person starts to talk (Row 4).

New Focus Targets Evaluation. To evaluate the set of New Focus Targets ((x, y, t)) generated from scene-specific LAAF, we compute the difference on focus plane in disparity units $\Delta d = \mathbb{D}(x, y) - \mathbb{D}(\hat{x}, \hat{y})$, and the temporal position offset in number of frames $\Delta t = t - \hat{t}$, where $(\hat{x}, \hat{y}, \hat{t})$ is the ground truth annotated New Focus Targets, using RVR-LAAF GUI.

5.3 Data-driven Video Autofocus

— CNN-Based Selection of New Focus Targets

To make autofocus fully automated on videos with any scene content, we present a first attempt at training a CNN (AF-Net) to predict New Focus Targets, replacing the scene-dependent detectors used in scene-specific LAAF. This is enabled using our RVR-LAAF GUI to annotate New Focus Targets on a large-scale video dataset. AF-Net takes in a sequence of frames centered at the query frame to predict the focus region (x, y) and the probability of the query frame being a New Focus Target.

³https://www.youtube.com/watch?v=tT_qv9ptauU&t=75s

⁴<http://dlib.net/>

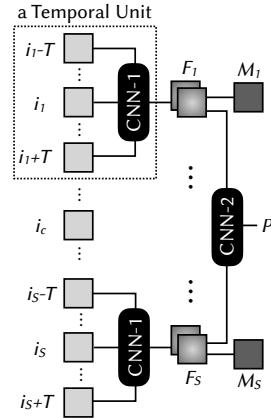


Fig. 6. AF-Net architecture predicts the focus region (M) for i_1, \dots, i_s and the probability (P) of the center frame i_c being a New Focus Target. The key of AF-Net is to cover a wide temporal range such that it sees past and future frames to make the prediction. In this illustration, CNN-1 has a temporal receptive field (f_T) of $2T$ and CNN-2 has a wide temporal receptive field of $2T \cdot s$, where s is the number of temporal units (f_s).

The key to LAAF is to analyze past and, particularly, future frames. A major challenge that arises is to have the network cover a wide temporal span of frames in a manner that is efficient in memory and computation. We introduce a temporal aggregation architecture consisting of two CNNs with different temporal receptive field sizes, as illustrated in Figure 6. The temporal receptive field is determined by the number of temporal units f_s , and the temporal coverage f_T inside each temporal unit. CNN-1 has a temporal receptive field of f_T to predict the focus region of the middle query frames (i_1, \dots, i_s). CNN-2 takes in feature maps generated from CNN-1, and predicts the probability of the global center frame (i_c), efficiently seeing a wider temporal receptive field of $f_s \cdot f_T$ frames. Loss of the network is a weighted sum of the evaluation metrics Δd and Δt we described in Section 5.2.

However, evaluating Δd requires disparity maps for all videos. We find that large-scale public video datasets that contain deep DOF are heavily-compressed [Abu-El-Haija et al. 2016; Wang et al. 2018a] and cannot be processed by our RGBD-HDR or have tracking be applied with high fidelity. We thus made 2 adjustments to training our AF-Net. First, instead of using $\Delta d = \mathbb{D}(x, y)$, we use its proxy (x, y) as supervision. Second, we choose to use a eye-tracking dataset DHF1K [Wang et al. 2018a] and a filtered version of its ground truth eye fixation map to supervise (x, y) , which we find to be highly correlated with focus region we annotated at New Focus Targets; supervision for t comes from our annotation. We call our annotated focus video dataset VAF.

We remove unsuitable videos (e.g. with jump cuts) and end up with 419 videos (640×360) of 20–60 seconds each in full frame rate (30 FPS). We reduce temporal sampling rate by a factor of 5 and set temporal stride of AF-Net to be 5. Both CNN-1 and CNN-2 architecture resembles ResNet-10, followed by a global average pooling for CNN-2. Training schedules are set to be the same as training our RGBD-HDR net.

6 RESULTS

We evaluate the key components of RVR (Section 6.1), and follow the metric described in Section 5.2 to quantitatively evaluate LAAF (Section 6.2). We also show a comparison against the autofocus system inside a high-end consumer camera Olympus EM1.2 in Section 6.4.

Aside from result figures, we would like to refer readers to the accompanying video for more visually distinguishable comparisons and results.

6.1 RVR Evaluation

We render shallow DOF video using the forward model in Equation 3 frame by frame, and compare rendering results with and without temporal coherency in Figure 3A. The focal planes are set to be the same for each comparison. We also show a forward model with and without occlusion-awareness in Figure 3D, and with and without HDR recovery in Figure 3B.

Evaluating intermediate network predictions on disparity and HDR maps against ground truth provides additional insight on rendering performance. For depth sensing, advanced range sensors such as LIDAR captures high-quality dense depth maps. Lightweight RGB-D cameras such as Intel RealSense [Keselman et al. 2017] have achieved adequate resolution for some consumer-grade applications, but still suffer from noisy output and limited precision around object edges. It still remains a challenge to obtain accurate depth for casual videos using portable devices. A survey on RGBD camera is written by Zollhöfer [Zollhöfer et al. 2018]. Current high-end smartphones such as iPhone X supports depth measurement using dual-pixels and dedicated post-processing to generate smooth, edge-preserving depth maps. A recent and relevant paper [Wang et al. 2018b] on monocular shallow DOF synthesis uses iPhone to construct the iPhone Depth Dataset for model training and testing. In a similar manner, we capture 50 test images using an iPhone X and extract disparity map as a proxy for ground truth to evaluate our predicted disparity map. We also apply a state-of-the-art monocular depth estimator, MegaDepth [Li and Snavely 2018], to these test images for comparison. We follow the quality metric proposed in [Scharstein and Szeliski 2002] to compute the RMS (root-mean-squared) error measured in disparity units between the predicted disparity map and its ground truth. For HDR evaluation, it is even more challenging to capture ground truth HDR using existing hardware sensors. We choose to use the public HDR test dataset constructed from exposure stacks from HDRCNN [Eilertsen et al. 2017a], and use their metric by computing the mean square error in the log space of the predicted linear image and its ground truth.

Figure 7A shows a histogram on disparity evaluation between our prediction and that from MegaDepth. Our indirect method without depth supervision produces comparable performance with MegaDepth that requires ground truth depth for training. In Figure 7B, we plot the histogram on HDR evaluation of the test images from [Eilertsen et al. 2017a]. It is expected that HDRCNN generates better quantitative performance as the model is trained with ground truth supervision, while our model is trained without direction supervision on HDR and on a different dataset. We provide an alternative way to recover HDR without ground truth required, and are able to produce HDR maps with adequate quality to render shallow DOF images (see Figure 3B). Importantly, per frame estimation of depth and HDR is not sufficient for rendering refocusable video, as we have shown that temporal coherence is also critical in Figure 3A and video at 01:46.

We design RVR as a flexible system to incorporate future works that improve upon disparity and HDR estimation for still images

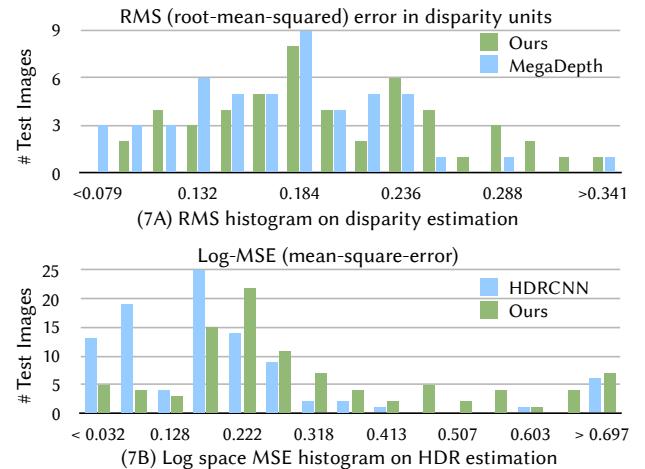


Fig. 7. Evaluation on predicted disparity and HDR against ground truth. Depth ground truth is obtained using dual-pixel on an iPhone X. HDR ground truth is from the public dataset described in [Eilertsen et al. 2017a].

and even for videos, or use camera sensors that support depth and HDR video streaming. A recent work [Wadhwa et al. 2018] uses dual-pixel imagery to estimate scene disparity for smartphone photography. When dual-pixel imagery is available, RVR could use its depth estimation as \mathbb{D} in the pipeline.

6.2 LAAF Evaluation

We evaluate the predicted New Focus Target (x, y, t) against ground truth $(\hat{x}, \hat{y}, \hat{t})$ (from GUI annotation) using Δd and Δt (See Section 5.2). For each test video, we compute the average focus depth difference $|\Delta d|$ across all frames, and the average temporal position difference $|\Delta t|$ across all New Focus Targets.

GUI-based Semi-Autofocus. We show an example in Figure 4C using RVR-LAAF GUI to annotate a New Focus Target (on the person's face). Figure 4C Row 3 presents that the user only needs to annotate the New Focus Target — selecting a focus region and creating a tracker, and the GUI will then track the selected region. The GUI also features fine tuning on the defocus strength and the focus puller duration to account for different story tone and visual sensitivity. High-res version of the GUI is shown in the accompanying video.

Evaluation on Action-aware Autofocus. We test LAAF on 11 casually collected videos with unexpected action that triggers focus depth change. We use action-aware LAAF with $\mathcal{H} = \{H_{\text{sal}}, H_{\text{act}}\}$, as described in Section 5.2 to compute New Focus Targets. Among all 11 test videos, 8 (72%) videos achieve $|\Delta t| < 15$, generating New Focus Targets that on average offset by less than half a second. Most test videos use LAAF to locate focus regions within 2 depth planes difference from ground truth. Quantitative results are shown in Table 1. A qualitative result is shown in Figure 4E, and more results at 05:00 in the video.

Evaluation on Data-driven Autofocus Detector. Among the 419 VAF videos, 40 videos are held out as test set for evaluation. To evaluate

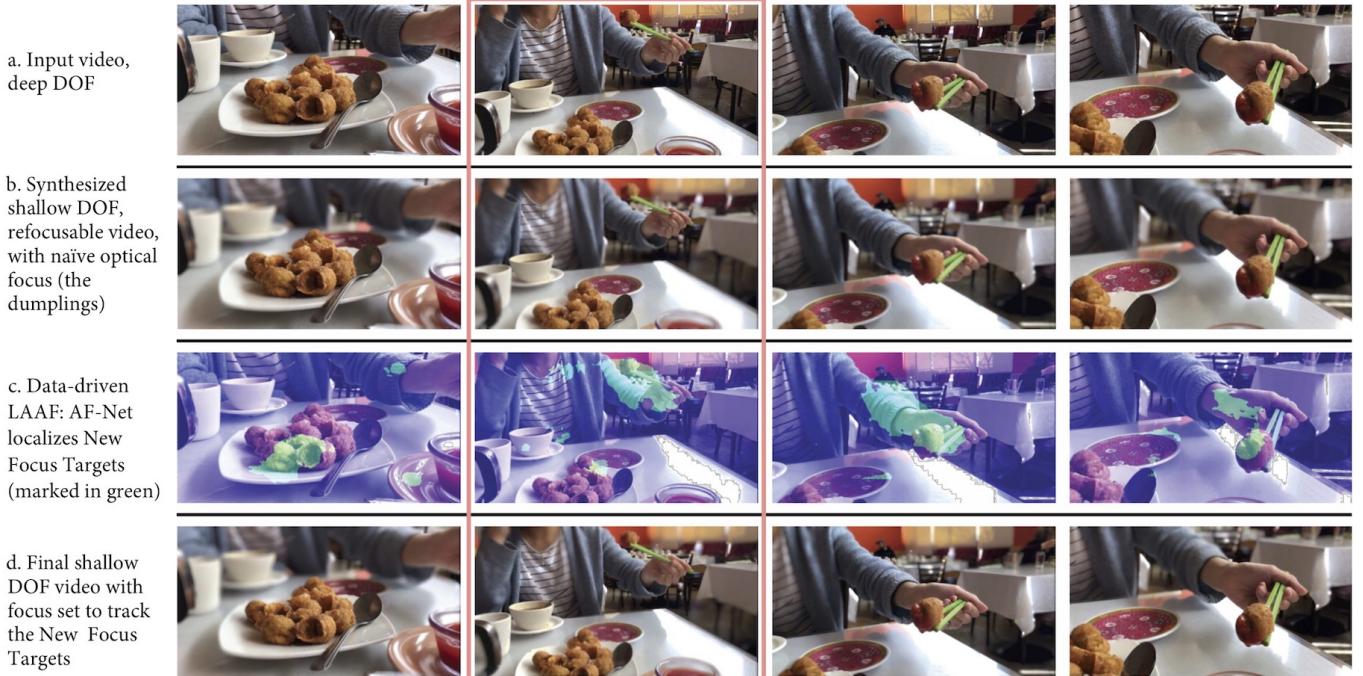


Fig. 8. Data-driven AF-Net to predict New Focus Targets for video autofocus. (Row 1) Input video shows a dumpling is picked up from the plate. (Row 2) naïve synthetic SDOF video would easily suffer from mis-focus when applied shallow DOF. (Row 3) Focus region prediction from AF-Net; green marks focus with higher probability. Second frame (bounded by red) is the predicted temporal position of the New Focus Target, when the person picks up the dumpling. (Row 4) The resulting video shows autofocus that tracks accurately on the dumpling. Video result can be found in the accompanying video.

Table 1. New Focus Target evaluation of action-aware LAAF on 11 casually collected videos using metrics introduced in Section 5.2. $|\Delta d|$ (in disparity units) evaluates the performance of computed depth plane over all frames for each video. $|\Delta t|$ (in number of frames) evaluates the average performance of temporal position of all New Focus Targets. Ground truth is obtained from hand-annotation using our RVR-LAAF GUI.

Video ID	1	2	3	4	5	6	7	8	9	10	11
$ \Delta d $	1.3	1.4	1.2	2.2	2.0	1.4	1.6	1.2	1.8	4.7	3.1
$ \Delta t $	6	8	3	15	7	10	14	20	7	> 30	16

New Focus Target, we run inference on a 100-frame clip to get the probability of each frame being the New Focus Target and select the frame with the highest probability. Note that for a test clip that does not have New Focus Targets (no focus depth changes required), we only evaluate Δd but not Δt . We also test on 30 collected smartphone videos, among which 11 are the action videos used to evaluate action-aware LAAF in Table 1. $|\Delta t|$ on the 11 action videos increase from 15.1 (using our specialized action-aware LAAF) to 21 (using AF-Net). However, AF-Net has the advantage of handling generic video contents and is able to achieve $|\Delta t|$ less than 1 second on all test videos (see Table 2). This indicates the potential of having a large-scale annotated video dataset with temporal aggregation (AF-Net) to tackle the challenging problem of video autofocus. One

Table 2. Results using AF-Net to predict New Focus Targets on VAF test and our collected videos. We compute the average $|\Delta t|$ across all test videos. $|\Delta d|$ is evaluated only on our test videos as we do not have disparity for VAF. We find AF-Net to perform slightly worse than action-aware LAAF on the action videos, but achieves reasonable performance on generic video contents, e.g. 1 second difference in t on all test videos. Ground truth for our collected test videos are from hand-annotation using RVR-LAAF GUI.

Dataset Method	VAF-test AF-Net	Ours-test AF-Net	Ours-action-test	
	Action-LAAF	AF-Net	1.99	2.21
$ \Delta d $	—	2.64		
$ \Delta t $	19.8	24.5	15.1	21.0

example result is shown in Figure 8. Video results are at 05:58 in the accompanying video.

Limitation of AF-Net. While AF-Net explores the potential of using machine learning for video autofocus, its requirement of large-scale video focus annotation is expensive and the dataset we accumulated was of modest size. Recently, unsupervised visual feature learned from large-scale unlabeled videos has shown to be effective for video understanding, such as object tracking via video colorization [Vondrick et al. 2018] and audio localization by learning to temporally align audio and video [Owens and Efros 2018]. We believe video autofocus can be likewise addressed by self-supervision

to learn from unlabeled internet-scale videos such as public movie clip dataset [Rohrbach et al. 2015].

6.3 Analysis of Artifacts in Output Video

The performance of our RVR-LAAF system is commensurate with a first prototype according to this new approach to the video auto-focus problem. For example, visible artifacts present in rendered videos due to imperfect disparity and HDR estimation. In this section, we classify these artifacts and discuss their causes by inspection and analysis of the real video results presented in previous sections, as well as experiments on a rendered video clip (see Supplementary Video B) that provides “ground truth” for comparison.

For the synthetic scene, we used a clip from the Blender Open Movie titled “Sintel” (2010), rendering this scene with a simulation of a small aperture similar to the deep DOF video captured by a modern smartphone camera. In Supplementary Video B, we show ablation results that compare ground truth disparity or HDR with estimations using our method and Eilertson *et al.* [2017b] for HDR and MegaDepth [Li and Snavely 2018] for disparity. It is important to note that there is a large domain shift between this synthetic rendering and the real training data used in all the estimation methods described, which disadvantages the estimated results. Nevertheless, the comparisons against ground truth provide empirical clues to support technical dissection of which errors in the system are associated with which classes of video artifacts.

The most visually prominent set of artifacts is due to errors in estimated disparity (see ablation comparisons in Supplementary Video B at 00:31). There are several classes of visual artifacts that may be seen. First, depth estimation across boundaries is imperfect, which is residual error in spite of our bilateral space processing. These edge errors cause prominent visual artifacts when in-focus regions are incorrectly blurred. Examples can be seen in the synthetic video result at 00:25 (ear) and in real video results at 05:13 (front person’s right shoulder). The second class of visual artifact is splotchiness of synthesized defocus blur, due to incorrect spatial variation of depth estimates on flat regions. This error tends to appear in regions at large disparity, and is residual error in spite of our triplet training procedure (Section 4.1) that helps to effectively reduce this problem. Examples can be seen in the synthetic video result at 00:31 (in the highlight region of the background arm), and in the real video result at 06:20 (background segments are incorrectly rendered sharper). The third class of visual artifact related to disparity estimation error is temporal fluctuation of the defocus blur, typically in regions of background. Examples of this error can be seen in the background of synthetic video results around 0:14, and in the real video results at 04:27 (behind conversation) and 05:20 (behind dog). This is residual temporal error in spite of compensation by our temporal stabilization module.

A second set of artifacts is related to errors in HDR estimation. Common examples of this error manifest as missing bokeh balls (false negative) in output video, while hallucinated bokeh balls (false positive) are generally rare. For example, real video results at 02:20 fail to recover HDR specular highlights on the glistening sea surface and are therefore missing expected salient bokeh balls. Synthetic video results at 00:43 underestimate the HDR value of the background figure’s arm and renders a darker defocused highlight.



Fig. 9. Cinematic bokeh rendering. Our system takes in the deep DOF image (on the left), and renders a cinematic bokeh using the predicted disparity and HDR map, and a lens shape that approximates the cinematography lens ARRI Master Anamorphic.

6.4 Compare against Market Camera

High-end consumer cameras with state-of-the-art autofocus technology still suffers from mis-focus, especially during a rapid subject change that requires focus to resolve accordingly. We capture a pair of videos of the same scene with a Olympus EM1.2 under f/2.8 and a smartphone. RVR-LAAF is then applied to the video from the smartphone to render shallow DOF with autofocus using our GUI. We compare the two videos and show RVR-LAAF tracks focus accurately while the DSLR fails to transition focus when making large subject change (See video 06:23).

An interesting application is to simulate defocus produced by expensive lenses. We apply our system to demonstrate a cinematic bokeh rendering in Figure 9 that approximates the ARRI Master Anamorphic⁵ lens for professional cinematography. Anamorphic lenses are prized by certain cinematographers, in part because of their ellipsoidal, decidedly non-circular, defocus blur (see Figure 9).

7 CONCLUSION

This paper introduces the problem of delivering cinema-like focus in casual videography (*i.e.* shallow DOF with context-aware focusing). We show that a traditional approach based on physical camera auto-focus is bound to fail, because errors in focus are baked into the video and focusing correctly in real-time requires error-prone guessing about where the action will go. We embrace this insight and take a fundamentally different approach with two parts: first, committing to rendering refocusable video from deep DOF video (RVR sub-system) rather than recording shallow DOF imagery; second, looking at future video frames to make focus decisions at every point in the video (LAAF sub-system) rather than only looking at past frames.

We built our RVR-LAAF prototype as a proof-of-concept for two main reasons. First, to show that we can achieve fundamentally better video auto-focus decisions by re-structuring the problem this way. And second, to show that it is tractable to attack the two sub-problems posed by this approach: synthesizing refocusable video and computing meaningful look-ahead autofocus decisions today. Regarding synthetic refocusable video, we summarized a broad array of technical approaches that the imaging and computational photography communities are actively advancing, from light field imaging to novel sensor designs to machine learning for depth inference. We are confident that performance and quality will improve rapidly.

⁵http://www.arduino.com/camera/cine_lenses/prime_lenses/anamorphic/

Regarding the problem of computing meaningful look-ahead autofocus decisions, we hope to have clearly conveyed the idea that this problem is also tractable and ripe for research. We believe that this area can also advance rapidly, given the broad range of current research in computer vision that can be brought to bear.

ACKNOWLEDGMENTS

This work is supported in part by NSF grant 1617794, an Alfred P. Sloan Foundation fellowship, and the Intel Faculty Support Program. Xuaner Zhang is supported by the J.K.Zee Fellowship Fund. We thank Alyosha Efros, Andrew Owens and anonymous reviewers for helpful discussions on the paper. We give a special thanks to people who kindly consent to be in the result videos.

REFERENCES

- Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).
- Jonathan T Barron, Andrew Adams, Yi-Chang Shih, and Carlos Hernández. 2015. Fast bilateral-space stereo for synthetic defocus. In *CVPR*.
- Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. 2016. Dynamic image networks for action recognition. In *CVPR*.
- Qifeng Chen and Vladlen Koltun. 2016. Full flow: Optical flow estimation by global optimization over regular grids. In *CVPR*.
- Qifeng Chen and Vladlen Koltun. 2017. Photographic image synthesis with cascaded refinement networks. In *ICCV*.
- Paul E Debevec and Jitendra Malik. 2008. Recovering high dynamic range radiance maps from photographs. *ACM Trans. on Graphics (TOG)* (2008).
- Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał Mantiuk, and Jonas Unger. 2017a. HDR image reconstruction from a single exposure using deep CNNs. *ACM Trans. on Graphics (TOG)* (2017).
- Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. 2017b. HDR image reconstruction from a single exposure using deep CNNs. *ACM Trans. on Graphics (TOG)* (2017).
- Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. 2018. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Trans. on Graphics (TOG)* (2018).
- G. D. Evangelidis and E. Z. Psarakis. 2008. Parametric Image Alignment Using Enhanced Correlation Coefficient Maximization. *PAMI* (2008).
- Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional two-stream network fusion for video action recognition. In *CVPR*.
- R Fontaine. 2017. A survey of enabling technologies in successful consumer digital imaging products. In *International Image Sensors workshop*.
- Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. 2016. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*.
- Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédéric Durand. 2017. Deep bilateral learning for real-time image enhancement. *ACM Trans. on Graphics (TOG)* (2017).
- Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. 2017. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*.
- Norman Goldberg. 1992. *Camera technology: the dark side of the lens*.
- Robert T Held, Emily A Cooper, James F O’A’brien, and Martin S Banks. 2010. Using blur to affect perceived distance and size. *ACM Trans. on Graphics (TOG)* (2010).
- João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. 2015. High-speed tracking with kernelized correlation filters. *PAMI* (2015).
- Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. 2017. Deeply supervised salient object detection with short connections. In *CVPR*.
- Eddy Ilg, Niklaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*.
- Aaron Isaksen, Leonard McMillan, and Steven J Gortler. 2000. Dynamically reparameterized light fields. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*.
- Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*.
- Neel Joshi and Larry Zitnick. 2014. *Micro-Baseline Stereo*. Technical Report.
- Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*.
- Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. 2017. Intel realsense stereoscopic depth cameras. In *CVPR Workshops*.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.
- Masahiro Kobayashi, Michiko Johnson, Yoichi Wada, Hiromasa Tsuboi, Hideaki Takada, Kenji Togo, Takafumi Kishi, Hidekazu Takahashi, Takeshi Ichikawa, and Shunsuke Inoue. 2016. A low noise and high sensitivity image sensor with imaging and phase-difference detection AF in all pixels. *ITE Trans. on Media Technology and Applications* (2016).
- Martin Kraus and Magnus Strengert. 2007. Depth-of-field rendering by pyramidal image processing. *CGF* (2007).
- Yevhen Kuznetsov, Jörg Stückler, and Bastian Leibe. 2017. Semi-supervised deep learning for monocular depth map prediction. In *CVPR*.
- Marc Levoy and Pat Hanrahan. 1996. Light Field Rendering. (1996).
- Marc Levoy and Yael Pritch. 2017. Portrait mode on the Pixel 2 and Pixel 2 XL smartphones.
- Zhengqi Li and Noah Snavely. 2018. MegaDepth: Learning Single-View Depth Prediction from Internet Photos. In *CVPR*.
- Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. 2017. Unsupervised video summarization with adversarial lstm networks. In *CVPR*.
- George Mather. 1996. Image Blur as a Pictorial Depth Cue. *Proc. Biological Sciences* (1996).
- Atsushi Morimitsu, Isao Hirota, Sozo Yokogawa, Isao Ohdaira, Masao Matsumura, Hiroaki Takahashi, Toshio Yamazaki, Hideki Oyaizu, Yalcin Incesu, Muhammad Atif, et al. 2015. *A 4M pixel full-PDAF CMOS image sensor with 1.58 μm 2× 1 On-Chip Micro-Split-Lens technology*. Technical Report.
- S. K. Nayar and Y. Nakagawa. 1994. Shape from focus. *PAMI* (1994).
- Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. 2005. *Light Field Photography with a Hand-held Plenoptic Camera*. Technical Report.
- Abhijit S Ogale, Cornelia Fermuller, and Yiannis Aloimonos. 2005. Motion segmentation using occlusions. *PAMI* (2005).
- Andrew Owens and Alexei A Efros. 2018. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. In *CVPR*.
- Jinsun Park, Yu-Wing Tai, Donghyeon Cho, and In So Kweon. 2017. A unified approach of multi-scale deep and hand-crafted features for defocus estimation. In *CVPR*.
- Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. 2017. Learning Features by Watching Objects Move. In *CVPR*.
- Alex Poms, Will Crichton, Pat Hanrahan, and Kayvon Fatahalian. 2018. Scanner: Efficient Video Analysis at Scale. (2018).
- Michael Potmesil and Indranil Chakravarty. 1982. Synthetic image generation with a lens and aperture camera model. *ACM Trans. on Graphics (TOG)* (1982).
- Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. 2016. Deepmatching: Hierarchical deformable dense matching. *IJCV* (2016).
- Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A Dataset for Movie Description. In *CVPR*.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*.
- Daniel Scharstein and Richard Szeliski. 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV* (2002).
- Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).
- Pratul P Srinivasan, Rahul Garg, Neal Wadhwa, Ren Ng, and Jonathan T Barron. 2018. Aperture Supervision for Monocular Depth Estimation. (2018).
- Meijun Sun, Ziqi Zhou, Qinghua Hu, Zheng Wang, and Jianmin Jiang. 2018. SG-FCN: A Motion and Memory-Based Deep Learning Model for Video Saliency Detection. *IEEE Trans. on Cybernetics* (2018).
- Jae young Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. 2012. Unstructured human activity detection from rgbd images. In *ICRA*.
- S. Suwajanakorn, C. Hernandez, and S. M. Seitz. 2015. Depth from focus with your mobile phone. In *CVPR*.
- Huixuan Tang, Scott Cohen, Brian L. Price, Stephen Schiller, and Kirilos N. Kutulakos. 2017. Depth from Defocus in the Wild. In *CVPR*.
- Michael W. Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi. 2013. Depth from Combining Defocus and Correspondence Using light-Field Cameras. In *ICCV*.
- Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. 2018. Tracking emerges by colorizing videos. In *ECCV*.
- Neal Wadhwa, Rahul Garg, David E. Jacobs, Bryan E. Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T. Barron, Yael Pritch, and Marc Levoy. 2018. Synthetic Depth-of-Field With A Single-Camera Mobile Phone. *ACM Trans. on Graphics (TOG)* (2018).
- Lijun Wang, Xiaohui Shen, Jianming Zhang, Oliver Wang, Zhe Lin, Chih-Yao Hsieh, Sarah Kong, and Huchuan Lu. 2018b. DeepLens: shallow depth of field from a single image. *ACM Trans. on Graphics (TOG)* (2018).

- Ting-Chun Wang, Jun-Yan Zhu, Nima Khademi Kalantari, Alexei A Efros, and Ravi Ramamoorthi. 2017. Light field video capture using a learning-based hybrid imaging system. *ACM Trans. on Graphics (TOG)* (2017).
- Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. 2018a. Revisiting Video Saliency: A Large-scale Benchmark and a New Model. In *CVPR*.
- Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy. 2005. High Performance Imaging Using Large Camera Arrays. (2005).
- Yang Yang, Haiting Lin, Zhan Yu, Sylvain Paris, and Jingyi Yu. 2016. Virtual DSLR: High Quality Dynamic Depth-of-Field Synthesis on Mobile Platforms. In *Digital Photography and Mobile Imaging*.
- Zhan Yu, Christopher Thorpe, Xuan Yu, Scott Grauer-Gray, Feng Li, and Jingyi Yu. 2011. Dynamic Depth of Field on Live Video Streams: A Stereo Solution. In *CCG*.
- Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video summarization with long short-term memory. In *ECCV*.
- Xuaner Zhang, Ren Ng, and Qifeng Chen. 2018. Single Image Reflection Removal with Perceptual Losses. In *CVPR*.
- Michael Zollhöfer, Patrick Stotko, Andreas Görilitz, Christian Theobalt, Matthias Nießner, Reinhard Klein, and Andreas Kolb. 2018. State of the Art on 3D Reconstruction with RGB-D Cameras. In *Computer Graphics Forum*.