# The "Vertigo Effect" on Your Smartphone: Dolly Zoom via Single Shot View Synthesis Supplement

Yangwen Liang Rohit Ranade Shuangquan Wang Dongwoon Bai
Jungwon Lee
Samsung Semiconductor Inc
{liang.yw, rohit.r7, shuangquan.w, dongwoon.bai, jungwon2.lee}@samsung.com

## 1. View Synthesis based on Camera Geometry

Recall from our paper, consider two pin–hole cameras A and B with camera centers at locations $\boldsymbol{C}_A$ and $\boldsymbol{C}_B$, respectively. From [2], based on the coordinate system of camera A, the projections of the same point $\boldsymbol{P} \in \mathbb{R}^3$ onto these two camera image planes have the following closed–form relationship

$$\begin{pmatrix} \boldsymbol{u}_B \\ 1 \end{pmatrix} = \frac{D_A}{D_B}\mathbf{K}_B\mathbf{R}(\mathbf{K}_A)^{-1}\begin{pmatrix} \boldsymbol{u}_A \\ 1 \end{pmatrix} + \frac{\mathbf{K}_B\boldsymbol{T}}{D_B} \quad (1)$$

where $\boldsymbol{T}$ can also be written as

$$\boldsymbol{T} = \mathbf{R}\left(\boldsymbol{C}_A - \boldsymbol{C}_B\right) . \quad (2)$$

Here, the $2 \times 1$ vector $\boldsymbol{u}_X$, the $3 \times 3$ matrix $\mathbf{K}_X$, and the scalar $D_X$ are the pixel coordinates on the image plane, the intrinsic parameters, and the depths of $\boldsymbol{P}$ for camera $X$, $X \in \{A, B\}$, respectively. The $3 \times 3$ matrix $\mathbf{R}$ and the $3 \times 1$ vector $\boldsymbol{T}$ are the relative rotation and translation of camera B with respect to camera A.

### 1.1. Generalized Single Camera System

Under dolly zoom condition, a generalized formula for camera movements on the horizontal and/or vertical directions along with the principal axis can be derived using Eq. 1. Let the camera center differences between camera 1 from position $\boldsymbol{C}_1^A$ to $\boldsymbol{C}_1^B$ be

$$\boldsymbol{C}_1^A - \boldsymbol{C}_1^B = \left(-m_1, -n_1, -t_1\right)^{\mathrm{T}} , \quad (3)$$

so that the camera moves by $m_1$, $n_1$ and $t_1$ in the horizontal, vertical and along the principal axis directions. We assume there is no relative rotation during camera translation. This assumption is valid due to the fact that we are creating a synthetic image at camera location $\boldsymbol{C}_1^B$ from image captured at location $\boldsymbol{C}_1^A$. As our paper shows, the intrinsic matrix $\mathbf{K}_1^B$ at camera center $\boldsymbol{C}_1^B$ is

$$\mathbf{K}_1^B = \mathbf{K}_1^A \mathrm{diag}\{k, k, 1\} , \quad (4)$$

where $k$ is the same as in the paper such that subject size on focus plane remains the same, i.e.

$$k = \frac{f_1^B}{f_1^A} = \frac{D_0 - t_1}{D_0} . \quad (5)$$

From Eq. 1, we can then obtain the closed–form solution for $\boldsymbol{u}_1^B$ in terms of $\boldsymbol{u}_1^A$ as:

$$\begin{aligned} \boldsymbol{u}_1^B = & \frac{D_1^A(D_0 - t_1)}{D_0(D_1^A - t_1)}\boldsymbol{u}_1^A + \frac{t_1(D_1^A - D_0)}{D_0(D_1^A - t_1)}\boldsymbol{u}_0 \\ & - \frac{(D_0 - t_1)f_1^A}{D_0(D_1^A - t_1)}\begin{pmatrix} m_1 \\ n_1 \end{pmatrix} . \end{aligned} \quad (6)$$

By setting $m_1 = n_1 = 0$ and $t_1 = t$, Eq. 6 reduces to Eq. (3) in our paper.

### 1.2. Generalized Dual Camera System

Similarly, a generalized formula for camera movements can be derived using Eq. 1 for this case as well. Here, we assume this dual camera system is well calibrated, and there is no relative rotation between cameras at location $\boldsymbol{C}_2$ and $\boldsymbol{C}_1^B$ (and $\boldsymbol{C}_1^A$). As in Section 1.1, let the camera center differences between camera 2 at location $\boldsymbol{C}_2$ and $\boldsymbol{C}_1^B$ be

$$\boldsymbol{C}_2 - \boldsymbol{C}_1^B = \left(-m_2, -n_2, -t_2\right)^{\mathrm{T}} , \quad (7)$$

so that the camera moves by $m_2$, $n_2$ and $t_2$ in the horizontal, vertical and along the principal axis directions. The baseline $b$ is assumed included in the $m_2$ and/or $n_2$. As in the paper, the intrinsic matrix $\mathbf{K}_2$ of camera 2 can be related to that of camera 1 at position $\boldsymbol{C}_1^A$ as

$$\mathbf{K}_2 = \mathbf{K}_1^A \mathrm{diag}\{k', k', 1\} \quad (8)$$

where the zooming factor $k'$ can be given as

$$k' = \frac{f_2}{f_1^A} = \frac{\tan(\theta_1^A/2)}{\tan(\theta_2/2)} . \quad (9)$$

From Eq. 1, we can obtain the closed–form solution for $\boldsymbol{u}_1^B$ in terms of $\boldsymbol{u}_2$ as

$$\boldsymbol{u}_1^B = \frac{D_2 k}{(D_2 - t_2)k'}\,(\boldsymbol{u}_2 - \boldsymbol{u}_0) + \boldsymbol{u}_0 + \frac{f_1^A k}{(D_2 - t_2)}\begin{pmatrix} m_2 \\ n_2 \end{pmatrix} \tag{10}$$

By setting $m_2 = b$, $n_2 = 0$ and $t_2 = t$, Eq. 10 reduces to Eq. (6) in our paper.

### 1.3. Shallow Depth of Field (SDoF)

Since the focus of this paper is not on the shallow depth of field (SDoF) rendering, any acceptable synthetic SDoF methods on a mobile device can be applied, cf. e.g. [4, 3, 1, 6]. However, to apply synthetic SDoF effect on each dolly zoom frame, the size of the circle of confusion (CoC) $c$ of the blur kernel is needed to be determined. Unlike synthetic SDoF effects for image capture, cf. e.g. [1], the blur strength varies not only according to the depth but also the dolly zoom translation parameter $t$ for the dolly zoom effect.

Assuming a thin lens camera model [2], the relation betwen $c$, lens aperture $A$, magnification factor $m$, depth to an object under focus $D_0$ and another object at depth $D$ can be given as [5]:

$$c = Am\frac{|D - D_0|}{D}\ , \tag{11}$$

where magnification factor $m$ is defined as

$$m = \frac{f}{D_0 - f}\ . \tag{12}$$

Eq. 11 and 12 are satisfied when there is no zooming applied for the camera, i.e. the focal length of the thin len $f$ is fixed [5]. Under the dolly zoom condition as introduced in Section 2 in our paper, the focal length changes according to the movement $t$ along the principal axis. Here, we denote the focal length with repect to $t$ as $f(t)$. Therefore, the relationship between $f(t)$ and $t$ is as shown in Section 2.1 in the paper, i.e.,

$$f(t) = \frac{D_0 - t}{D_0} f(0)\ . \tag{13}$$

Accordingly, the magnification factor $m(t)$ with respect to $t$ can be obtained as

$$m(t) = \frac{f(t)}{(D_0 - t) - f(t)}\ . \tag{14}$$

By substituting Eq. 13 into Eq. 14, we can obtain

$$m(t) = \frac{f(0)}{D_0 - f(0)} = m(0) = m\ . \tag{15}$$

Eq. 15 perfectly aligns with pinhole camera model under dolly zoom, i.e. the magnification factor for subjects in focus is fixed. Also, the relative depth $|D - D_0|$ between subjects within the scene remains constant for single image capture. Assuming the lens aperture $A$ remains the same, we can obtain the CoC size as:

$$c(t) = Am\frac{|D - D_0|}{D - t} = c(0)\frac{D}{D - t}\ , \tag{16}$$

where $c(t)$ is the CoC diameter for an object at depth $D$ and the camera translation $t$ along the principal axis. As we can observe, the initial CoC $c(0)$ is not of concern in our derivation, and any flavor of initial CoC size or shape can be applied, cf. e.g. [3, 6]. Eq. 16 provides the CoC size update to mimic the real dolly zoom effect.

## 2. Experiment Results

### 2.1. Supplementary Videos

We have included supplementary videos showing the results of our synthesis pipeline described in Section 2.6 of the main paper. In the dual camera synthesis case, the video shows the input images ($\mathbf{I}_1$ and $\mathbf{I}_2$), their depth maps ($\mathbf{D}_1$ and $\mathbf{D}_2$), the synthesized views and the corresponding ground truth for successively greater dolly zoom angles (without and with the SDoF effect) compiled into a video sequence. Similarly, the single camera synthesis case shows the input image ($\mathbf{I}$) and its depth map ($\mathbf{D}$), the generation of the two images ($\mathbf{I}_1$ and $\mathbf{I}_2$) and their depth maps ($\mathbf{D}_1$ and $\mathbf{D}_2$) (formed from image $\mathbf{I}$ and depth $\mathbf{D}$ of each image set as described in Section 2.6, Steps 2–3 of the main paper) and the results of our synthesis pipeline for successively greater dolly zoom angles. This case also shows the result of our pipeline applied to an image set from the smartphone dataset (without and with the SDoF effect).

### 2.2. Qualitative Results

Figures 1, 2 and 3 show the results of our method applied to images from the smartphone dataset. In each example, (a) and (b) are the input images $\mathbf{I}_1$ and $\mathbf{I}_2$ (formed from image $\mathbf{I}$ of each image set as described in Section 2.6, Steps 2–3 of the main paper) and (c) is the input image $\mathbf{I}_1$ with the SDoF effect applied. We then apply the single camera single shot view synthesis pipeline described in Section 2.6 (Steps 4–8) of the main paper to generate synthesized images for successfully greater dolly zoom angles ((d)–(i)).

## References


[1] Jonathan T. Barron, Andrew Adams, YiChang Shih, and Carlos Hernández. Fast bilateral–space stereo for synthetic defocus. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[2] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Wiley, 2007.

[3] David E. Jacobs, Jongmin Baek, and Marc Levoy. Focal stack compositing for depth of field control. *Stanford Computer Graphics Laboratory Technical Report 2012-1*, October 2012.

[4] M. Kraus and M. Strengert. Depth-of-field rendering by pyramidal image processing, 2007.

[5] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer, 2011.

[6] 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. Synthetic depth–of–field with a single–camera mobile phone. *ACM Transactions on Graphics (TOG)*, 37(4):64:1–64:13, July 2018.

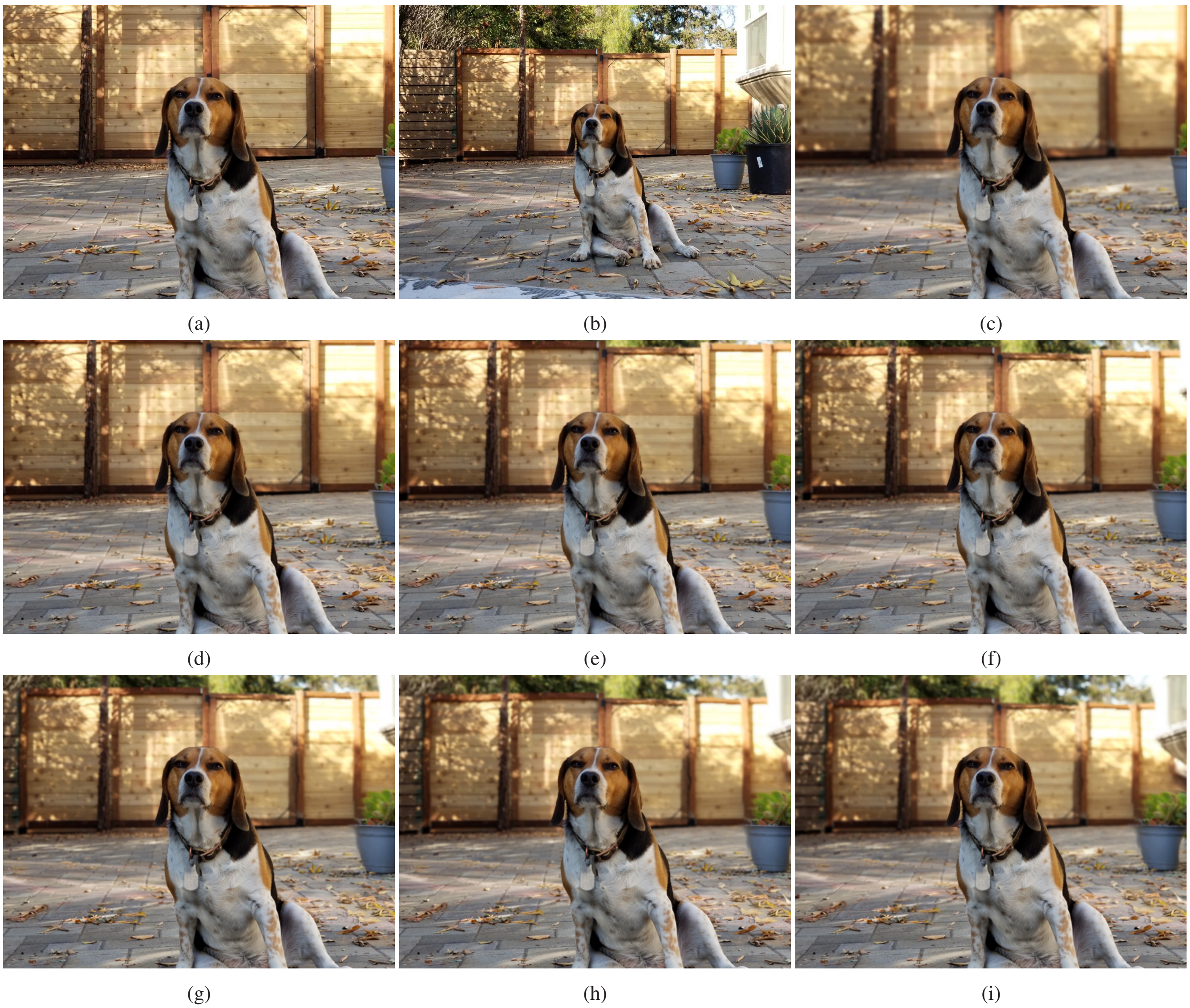


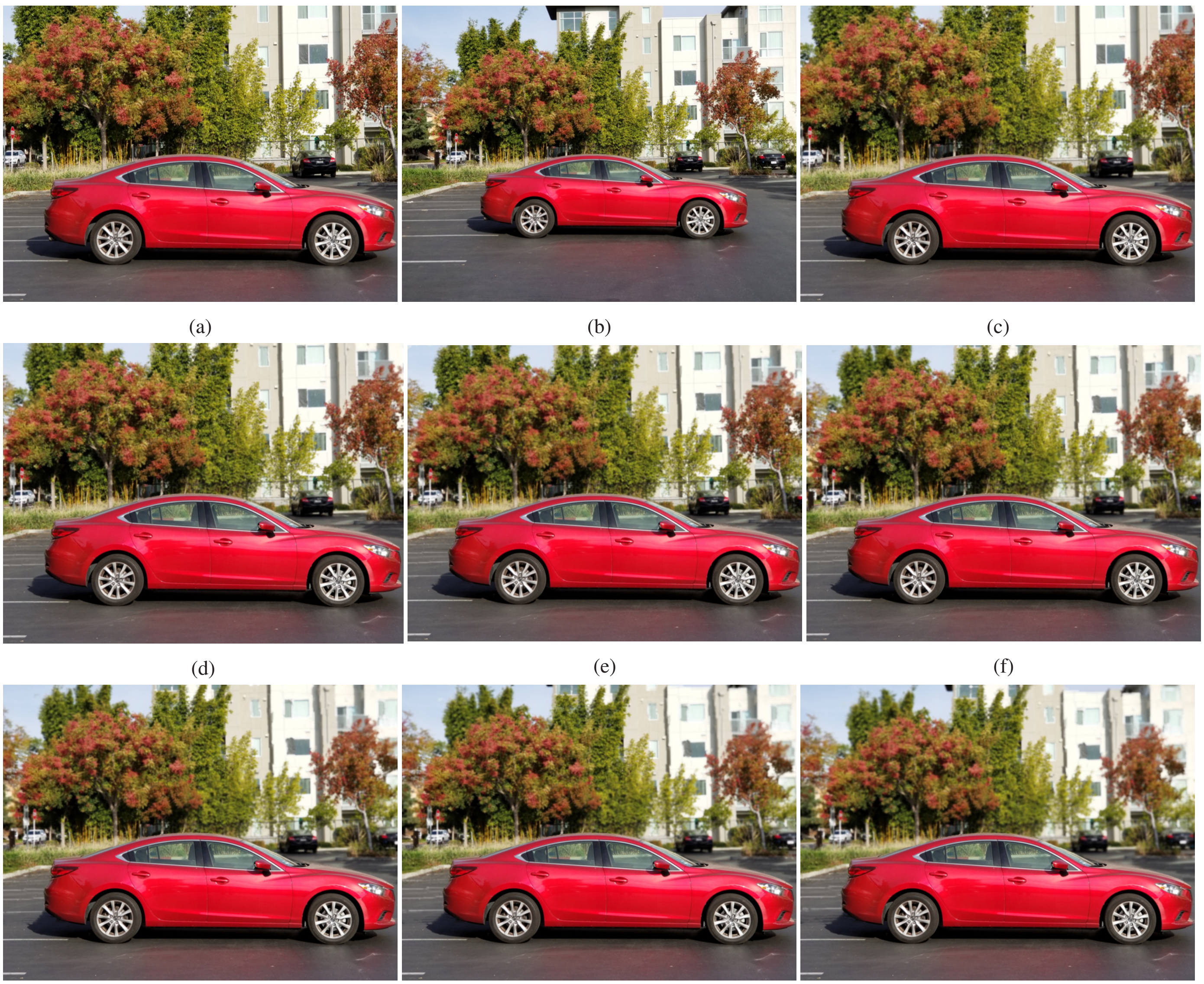

Figure 1: Single camera single shot dolly zoom view synthesis with smartphone dataset - Example 1. Here (a) and (b) are the input images $\mathbf{I}_1$ and $\mathbf{I}_2$, (c) is the image $\mathbf{I}_1$ from (a) with the SDoF effect applied, while (d)–(i) are the synthesized images with our method (after the application of the SDoF effect), for successively greater dolly zoom angles.

(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 2: Single camera single shot dolly zoom view synthesis with smartphone dataset - Example 2. Here (a) and (b) are the input images $\mathbf{I}_1$ and $\mathbf{I}_2$, (c) is the image $\mathbf{I}_1$ from (a) with the SDoF effect applied, while (d)–(i) are the synthesized images with our method (after the application of the SDoF effect), for successively greater dolly zoom angles.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3: Single camera single shot dolly zoom view synthesis with smartphone dataset - Example 3. Here (a) and (b) are the input images $\mathbf{I}_1$ and $\mathbf{I}_2$, (c) is the image $\mathbf{I}_1$ from (a) with the SDoF effect applied, while (d)–(i) are the synthesized images with our method (after the application of the SDoF effect), for successively greater dolly zoom angles.