## Auxiliary material 2: vibration affects the mathematical derivation process

### 1 Vibration mechanism of breaking FV

FV is a key concept in SFF, serving as the core data carrier for 3D reconstruction technology based on multi-focus image sequences. The traditional method assumes that the image sequence is strictly aligned, that the distribution of pixel clarity with focal length follows an ideal Gaussian distribution, and that high-precision depth estimation is achieved by Gaussian fitting of the focus signal or peak detection. However, in practice, environmental vibration during image acquisition will distort the distribution characteristics of the focus curve, and the traditional depth estimation method will fail, leading to a significant increase in depth estimation error.

#### 1.1 Spatial Aliasing

The translational vibration occurring in the x-y plane during image acquisition is termed the spatial aliasing effect. Assuming the actual depth at a point $P(x_0, y_0)$ in the image sequence is $z=\mu$, the distribution characteristics of the idealized focus signal for this point satisfy the following mathematical model:

$$D_{ideal}(z) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(z-\mu)^2}{2\sigma^2}} \tag{1}$$

Where $\sigma$ is determined by the depth of field (DoF) of the optical system, under ideal vibration-free conditions, during the acquisition of the multi-focus image sequence, the translation stage carrying the sample moves along the optical axis in steps of a preset size, this method ensures that the position of point $P(x_0, y_0)$ remains fixed in every image corresponding to each sampling point. $z_k$.

However, in practice, during the acquisition of the $k$-th image layer, in-plane x-y offsets within the o-xoy plane cause the projected coordinates of point P to undergo random displacements ( $\Delta x_k$, $\Delta y_k$). Consequently, the sharpness value recorded at position ($x_0$, $y_0$) in the Focus Volume (FV) actually originates from a neighboring point $P'$ with coordinates ($x_0+\Delta x_k$, $y_0+\Delta y_k$). According to the Central Limit Theorem, if the vibration results from the superposition of a large number of independent, minute random disturbances, its aggregate effect tends to approximate a Gaussian distribution. Therefore, it is assumed that the random in-plane offsets $\Delta x_k$, $\Delta y_k$ follow a Gaussian distribution with a mean of zero:

$$\Delta x_k \sim \mathcal{N}\ (0,\ \sigma_x^2)\ ,\quad \Delta y_k \sim \mathcal{N}\ (0,\ \sigma_y^2) \tag{2}$$

When performing depth estimation at positions where the object surface depth varies continuously, in-plane vibration causes the sharpness value at each sampling point to originate from a different spatial location. Taking point $P$ as an example, its sharpness value in the FV is essentially a mixture of sharpness values from all possible offset positions within its neighborhood. Specifically, this manifests as the pseudo-sharpness value of point $P$ resulting from a continuous translation and mixing of sharpness values along a Gaussian curve. Assuming the point spread function (PSF) of the image is h($x$, $y$), the recorded pseudo-sharpness value for point $P$ in the FV is given by:

$$D_{obs}(z_k) = \iint h(\Delta x, \Delta y) \cdot D(z_k; \mu + \delta z(\Delta x, \Delta y))d\Delta x d\Delta y \tag{3}$$

Where $\delta z(\Delta x, \Delta y)$ represents the equivalent depth deviation induced by the in-plane vibration shift, which is determined by the slope of the object surface.

In such cases, in-plane vibration can cause the focus signal to appear as a broadened, shifted, or complex hybrid peak. When performing depth estimation at surface points where depth discontinuities occur along the in-plane direction, discrete depth jumps on the object surface, combined with in-plane vibration, result in a multi-peak response in the sharpness values. Specifically, the pseudo-sharpness value at a given point $P$ is obtained from the discrete superposition of sharpness contributions originating from points at different depths. If the actual depth at the nominal position P is $z=\mu$, and the depth at the shifted position $P'$ (due to vibration) is $z=\mu'$, then the observed signal can be expressed as:

$$D_{obs}(z_k) = \omega_1 D(z;\ \mu_1,\ \sigma) + \omega_2 D(z;\ \mu_2,\ \sigma) \tag{4}$$

Where the vibration offset and the PSF determine the weights $\omega_1$ and $\omega_2$, under such conditions, in-plane vibration causes the focus curve to exhibit either a double-peak or a flattened peak shape.

## 1.2 Sampling Distortion

The vibration along the Z-axis during image acquisition is referred to as the sampling-point deviation phenomenon. As derived from the image acquisition principle, under vibration-free conditions, the discrete positions of the focal sampling points along the Z-axis satisfy a linear equation:

$$z_k = z_0 + k\ \Delta \mathrm{z} \tag{5}$$

Where $z_0$ is the initial height, k is the layer index, and $\Delta z$ is the acquisition step size of the vision system. After introducing a random disturbance $\delta z_k \sim \mathcal{N}(0，\ \sigma_z^2)$, the actual sampling position becomes:

$$z_k^{'} = z_k + \delta_{z_k} \tag{6}$$

At this point, the observed value of clarity becomes:

$$D_{obs}(z_k^{'}) = D_{ideal}(z_k + \delta z_k;\ \mu, \sigma) \tag{7}$$

Assuming the actual peak position is $\mu$, to ensure the computed estimated value more closely approximates the proper depth, we solve for the optimal solution. $\hat{\mu}$ Of the following expression via maximum likelihood estimation (MLE):

$$\mathcal{L}(\mu) = \prod_k p\big(D_{obs}(z_k^{'})\big|\mu\big) \tag{8}$$

The Taylor series expansion of $D_{obs}(z_k^{'}$ ) about zk yields the following expression:

$$D_{obs}(z_k^{'}\ ) \approx D(z_k) + \delta z_k \cdot \dot{D}(z_k) + \tfrac{1}{2}\delta {z_k}^2 \cdot \ddot{D}(z_k) \tag{9}$$

Where, $\dot{D}$ and $\ddot{D}$ Represent the first and second derivatives of the Gaussian function $D(z_k)$, respectively. It is known that the maximum likelihood condition requires:

$$\sum_k \frac{\partial\ lnD(z_k^{'};\mu)}{\partial \mu} = 0 \tag{10}$$

By incorporating the maximum likelihood condition into the expansion, the vibration-induced estimation bias term can be derived:

$$\hat{\mu} - \mu \approx \frac{\sum \delta z_k \cdot \dot{D}(z_k)}{\sum \ddot{D}(z_k)} \tag{11}$$

The above equation indicates that vibration disturbance linearly amplifies the error through the gradient. $\dot{D}(z_k)$ of the sharpness curve. Under ideal conditions without out-of-plane vibration, the first derivative $\dot{D}(z_k)$ of the Gaussian function $D(z_k)$ exhibits odd symmetry near its peak, causing positive and negative deviations to cancel out and resulting in an expected bias of zero. However, when actual disturbances introduce the random variable $\delta z_k$ The expected value of the bias is not necessarily zero. Particularly when the vibration amplitude is comparable to the depth of field, the systematic error becomes pronounced.

1.3 Joint Coupling Effect

The coupled effect arising from the combined action of translational vibration in the o-xoy plane and vertical vibration in the o-yoz plane causes the construction process of the FV to degenerate into a non-stationary stochastic process. Let the vibration-induced displacement offsets along the three axes be denoted as $(\Delta x_k, \Delta y_k, \delta z_k)$. According to random vibration theory, the response of a multi-degree-of-freedom system is appropriately described by a covariance matrix to characterize modal coupling. Therefore, the three-axis displacement offsets follow a three-dimensional joint Gaussian distribution:

$$\Delta_k \sim \mathcal{N}\text{（}0\text{，}\Sigma\text{）} \quad , \quad \Sigma = \begin{bmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y & \rho_{xz}\sigma_x\sigma_z \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 & \rho_{yz}\sigma_y\sigma_z \\ \rho_{xz}\sigma_x\sigma_z & \rho_{yz}\sigma_y\sigma_z & \sigma_z^2 \end{bmatrix} \tag{12}$$

Where, $\sigma_x^2$, $\sigma_y^2$, and $\sigma_z^2$ represent the vibration variances along the x, y, and z directions, respectively. $\rho_{xy}$, $\rho_{xz}$, and $\rho_{yz}$ Denote the correlation coefficients between the three directions, with their values ranging from -1 to 1.

Consequently, the observed focus curve can be expressed as the convolution of the ideal curve with the vibration distribution:

$$D_{obs}(z) = \iiint D(x+\Delta x,\ y+\Delta y,\ z+\delta z)\cdot p(\Delta x,\ \Delta y,\ \delta z)\,d\Delta x\ d\Delta y\ d\delta z \tag{13}$$

Where, $p(\Delta x, \Delta y, \delta z)$ is the probability density function of the vibration. Can approximate the mean squared error (MSE) of the depth estimation as:

$$\mathbb{E}[(\hat{\mu}-\mu)^2] \approx \frac{\sigma_z^2\cdot\int(\dot{D}(z))^2\,dz + \sigma_x^2\sigma_y^2\cdot\int(\nabla_{xy}D(z))^2\,dz}{(\int\ddot{D}(z)dz)^2} \tag{14}$$

The error $\epsilon$ between the ground truth depth value and the estimated depth value is finally obtained, which satisfies:

$$\mathbb{E}[\epsilon^2] \propto \sigma^2_{\text{振动}}\cdot\ \left(\frac{\partial^2 D}{\partial z^2}\right)^{-1} \tag{15}$$

As indicated by the equation above, the magnitude of the error is jointly determined by the vibrational energy and the sensitivity of the optical system.

In summary, in-plane vibration leads to a mismatch in the spatial coordinate system of the imaging field, thereby disrupting the spatial consistency of the FV. Out-of-plane vibration introduces phase errors, disrupting the temporal stability of focal sampling. These two perturbations create a convolutional coupling in the spatiotemporal domain. When superimposed with the nonlinear distortion of the gradient field, they cause the focus signal to exhibit distinctly non-Gaussian distribution characteristics. Consequently, traditional algorithms based on single-peak detection or curve fitting fail under these conditions. This situation urgently requires the development of a multidimensional correction framework to reconstruct high-fidelity FV curves from distorted data.