MODELING THE IMPACT OF KEYPOINT DETECTION ERRORS ON LOCAL DESCRIPTOR SIMILARITY

SUPPLEMENTAL MATERIAL

André Araujo, Haricharan Lakshman, Roland Angst, Bernd Girod

Department of Electrical Engineering, Stanford University, CA

ABSTRACT

We provide further derivations and experimental results. First, we present detailed derivations to obtain approximate closed-form expressions for the expected squared L_2 distance when translation errors are known, or when they are uniformly distributed. Second, we provide more experimental results to validate the Gamma distribution model derived for component-wise squared L_2 distances.

Appendix A: Supplemental derivations

In this section, we first present a useful approximation of $\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})$, which works well in practice. With this expression, we can write $E[\|f_A - f_B\|_2^2|\boldsymbol{\Delta}\mathbf{v}]$ in closed-form. Then, we present the derivation of the expected values of $\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})$ and $\mathrm{Pyr}(\boldsymbol{\Delta}\mathbf{v})$, using a uniform distribution of translation errors. This allows us to obtain a closed-form expression for $E[\|f_A - f_B\|_2^2]$.

Approximation of $\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})$

$\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})$ denotes the number of pixel pairs within the non-overlap region of a given spatial bin which are separated by an $[i, j]$ displacement, when the spatial bin is shifted by $\boldsymbol{\Delta}\mathbf{v}$ with respect to the ground-truth spatial bin. Note that we are not interested in $\mathrm{N}_{0,0}(\boldsymbol{\Delta}\mathbf{v})$, as the summations that use $\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})$ explicitly discard the $[0, 0]$ displacement. This happens because the $[0, 0]$ displacement gives rise to the variance of the random variable under consideration, which can be calculated in closed-form in our model.

Fig. 1 gives three examples of pixel grids with some displacements, with $w = h = 4$. The blue grid corresponds to the grid in the correct position, while the red one corresponds to the grid with the incorrect keypoint detection. To give some examples of the values we want to compute, the case (a) has:

- $\mathrm{N}_{1,0}([1, -1]) = \mathrm{N}_{-1,0}([1, -1]) = 3$
- $\mathrm{N}_{0,1}([1, -1]) = \mathrm{N}_{0,-1}([1, -1]) = 3$
- $\mathrm{N}_{1,1}([1, -1]) = \mathrm{N}_{-1,-1}([1, -1]) = 1$
- $\mathrm{N}_{2,2}([1, -1]) = \mathrm{N}_{-2,-2}([1, -1]) = 1$
- $\mathrm{N}_{3,3}([1, -1]) = \mathrm{N}_{-3,-3}([1, -1]) = 1$

Clearly, we see that $\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v}) = \mathrm{N}_{-i,-j}(\boldsymbol{\Delta}\mathbf{v})$, so we only need to compute $\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})$ for half of all possible $[i, j]$.

It is difficult to express $\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})$ exactly as a function of i, j and $\boldsymbol{\Delta}\mathbf{v}$, so we use an approximation. We divide the non-overlap region into two regions, as in Fig. 2: (i) a "vertical" region (green), which is generated by horizontal shifts and (ii) a "horizontal" region (orange), which is generated by vertical shifts. We calculate $\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})$ for each of these regions, and subtract the contribution from their intersection:

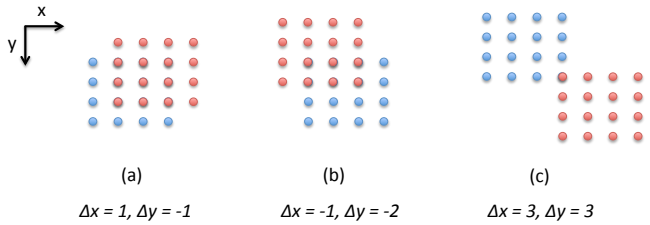

Fig. 1: Three different examples of pixel grids with $w = h = 4$. The samples in blue represent the grid in the correct position, while the red ones represent the grid used for the incorrect keypoint detection.

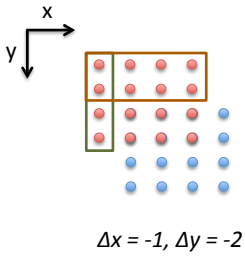

Fig. 2: Example of pixel grids with $w = h = 4$, showing the two regions into which the non-overlap area is divided: vertical (green) and horizontal (orange). The samples in blue represent the grid in the correct position, while the red ones represent the grid used for the incorrect keypoint detection.

$$\begin{aligned}\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v}) \approx\ & \max(0, (h - |j|))\max(0, |\Delta x| - |i|) \\ & + \max(0, (w - |i|))\max(0, |\Delta y| - |j|) \\ & - \max(0, |\Delta x| - |i|)\max(0, |\Delta y| - |j|) \end{aligned} \quad (1)$$

Considering Fig. 2, we can see that this expression is an approximation since we are not taking into account the pairs formed by, say, a pixel on the right of the orange region and a pixel on the bottom of the green region. More concretely, $\mathrm{N}_{-3,3}([-1, -2]) = 1$, but this approximation gives $\mathrm{N}_{-3,3}([-1, -2]) = 0$. However, this expression works well because the pixel pairs which are not taken into consideration are usually the ones which are the most distant, and the covariance between distant pixels is weaker. In this example, our approximation gives $\mathrm{N}_{0,1}([-1, -2]) = 6$, $\mathrm{N}_{1,1}([-1, -2]) = 3$, $\mathrm{N}_{2,0}([-1, -2]) = 4$ $\mathrm{N}_{2,1}([-1, -2]) = 2$, which are all correct.

We can then use (1) to compute the final closed-form expression for $E[\|f_A - f_B\|_2^2|\boldsymbol{\Delta}\mathbf{v}]$.

Expected value of $\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})$

We consider a discrete uniform distribution of translation errors. In this case, the distribution is separable with probability mass function of $\frac{1}{U^2}$ at each point within $[-\frac{U}{2}, \frac{U}{2}-1] \times [-\frac{U}{2}, \frac{U}{2}-1]$, with $U > 0$ and multiple of 2.

Using the approximation (1), we obtain:

$$\begin{aligned} & E_{\boldsymbol{\Delta}\mathbf{v}}[\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})] \\ & \approx \max(0,(h-|j|))E_{\Delta x}[\max(0,|\Delta x|-|i|)] \\ & + \max(0,(w-|i|))E_{\Delta y}[\max(0,|\Delta y|-|j|)] \\ & - E_{\Delta x}[\max(0,|\Delta x|-|i|)]E_{\Delta y}[\max(0,|\Delta y|-|j|)] \end{aligned} \tag{2}$$

The calculation of $E_{\boldsymbol{\Delta}\mathbf{v}}[\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})]$ thus requires the computation of $E_{\Delta x}[\max(0,|\Delta x|-|i|)]$. Using the discrete uniform distribution:

$$\begin{aligned} & E_{\Delta x}[\max(0,|\Delta x|-|i|)] = \\ & = \frac{1}{U}\sum_{\Delta x=-\frac{U}{2}}^{\frac{U}{2}-1}\max(0,|\Delta x|-|i|) \\ & = \frac{1}{U}\left[\max\left(0,\frac{U}{2}-|i|\right)+2\times\sum_{\Delta x=0}^{\frac{U}{2}-1}\max(0,|\Delta x|-|i|)\right] \end{aligned} \tag{3}$$

In the case where $|i| \leq \frac{U}{2}-1$, we can derive:

$$\begin{aligned} \sum_{\Delta x=0}^{\frac{U}{2}-1}\max(0,|\Delta x|-|i|) &= \sum_{\Delta x=|i|}^{\frac{U}{2}-1}(|\Delta x|-|i|) \\ &= \frac{\left(\frac{U}{2}-|i|\right)\left(\frac{U}{2}-1-|i|\right)}{2} \end{aligned} \tag{4}$$

In the case where $|i| > \frac{U}{2}-1$, clearly $\sum_{\Delta x=0}^{\frac{U}{2}-1}\max(0,|\Delta x|-|i|) = 0$. Finally, we obtain:

$$\begin{aligned} & E_{\Delta x}[\max(0,|\Delta x|-|i|)] = \\ & = \frac{1}{U}\left[\max\left(0,\frac{U}{2}-|i|\right)+\left(\frac{U}{2}-|i|\right)\max\left(0,\frac{U}{2}-1-|i|\right)\right] \end{aligned} \tag{5}$$

The final expression for $E_{\Delta y}[\max(0,|\Delta y|-|j|)]$ is very similar, naturally. Thus, we can easily compute $E_{\boldsymbol{\Delta}\mathbf{v}}[\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})]$ by using (2).

Expected value of $\mathrm{Pyr}(\boldsymbol{\Delta}\mathbf{v})$

Consider the same discrete uniform distribution of translation errors as before: a separable distribution with probability mass function of $\frac{1}{U^2}$ at each point within $[-\frac{U}{2}, \frac{U}{2}-1] \times [-\frac{U}{2}, \frac{U}{2}-1]$, with $U > 0$ and multiple of 2. In this case:

$$\begin{aligned} & E_{\boldsymbol{\Delta}\mathbf{v}}[\mathrm{Pyr}(\boldsymbol{\Delta}\mathbf{v})] \\ & = \frac{1}{U^2}\sum_{\Delta x=-\frac{U}{2}}^{\frac{U}{2}-1}\sum_{\Delta y=-\frac{U}{2}}^{\frac{U}{2}-1}\mathrm{Pyr}(\Delta x,\Delta y) \\ & = \frac{1}{U^2}\sum_{\Delta x=-\frac{U}{2}}^{\frac{U}{2}-1}\sum_{\Delta y=-\frac{U}{2}}^{\frac{U}{2}-1}\frac{(w-|\Delta x|)(h-|\Delta y|)}{w^2h^2} \\ & = \frac{1}{U^2w^2h^2}\sum_{\Delta x=-\frac{U}{2}}^{\frac{U}{2}-1}(w-|\Delta x|)\sum_{\Delta y=-\frac{U}{2}}^{\frac{U}{2}-1}(h-|\Delta y|) \\ & = \frac{1}{wh}\left[1-\frac{U(w+h)}{4wh}+\frac{U^2}{16wh}\right] \end{aligned} \tag{6}$$

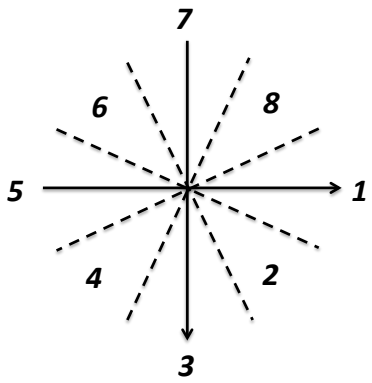

Fig. 3: Gradient orientation bin (d) numbering convention used in our work.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Fig. 4: Spatial bin (n) numbering convention used in our work.

where the last step is obtained by using the fact that $\sum_{k=0}^{K} = \frac{K(K+1)}{2}$, and standard mathematical derivations. In the case of $w = h$, we can further simplify (6) to:

$$E_{\boldsymbol{\Delta}\mathbf{v}}[\mathrm{Pyr}(\boldsymbol{\Delta}\mathbf{v})] = \frac{1}{w^2}\left(1-\frac{U}{4w}\right)^2 \tag{7}$$

Together with the expression for $E_{\boldsymbol{\Delta}\mathbf{v}}[\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})]$ derived in the previous subsection, the expression for $E_{\boldsymbol{\Delta}\mathbf{v}}[\mathrm{Pyr}(\boldsymbol{\Delta}\mathbf{v})]$ can then be used to derive the final closed-form expression for $E[\|f_A - f_B\|_2^2]$.

Appendix B: Supplemental experimental results

In this section, we present further experimental results for the distribution of $z_{d,n}^2$ given $\boldsymbol{\Delta}\mathbf{v}$. Our objective is to provide further evidence that the Gamma distribution approximation for $z_{d,n}^2$, using M-S, works well. To clarify which spatial and orientation bins are used, Fig. 3 and Fig. 4 present the conventions we use for numbering them. Fig. 5 and Fig. 6 present estimated distributions for different orientation and spatial bins, using both the SMVS and the CNN2h datasets, plotted against ground-truth distributions measured from data. These figures show the estimated distributions using M-IID and M-S, plotted for $\boldsymbol{\Delta}\mathbf{v}_1$, $\boldsymbol{\Delta}\mathbf{v}_2$ and $\boldsymbol{\Delta}\mathbf{v}_3$ (as described in the experimental section of the main part of this paper). We can infer that the M-IID models (in green) estimate the distributions poorly. The M-S model (in red) estimates the ground-truth distributions much better, certainly capturing the main trends.

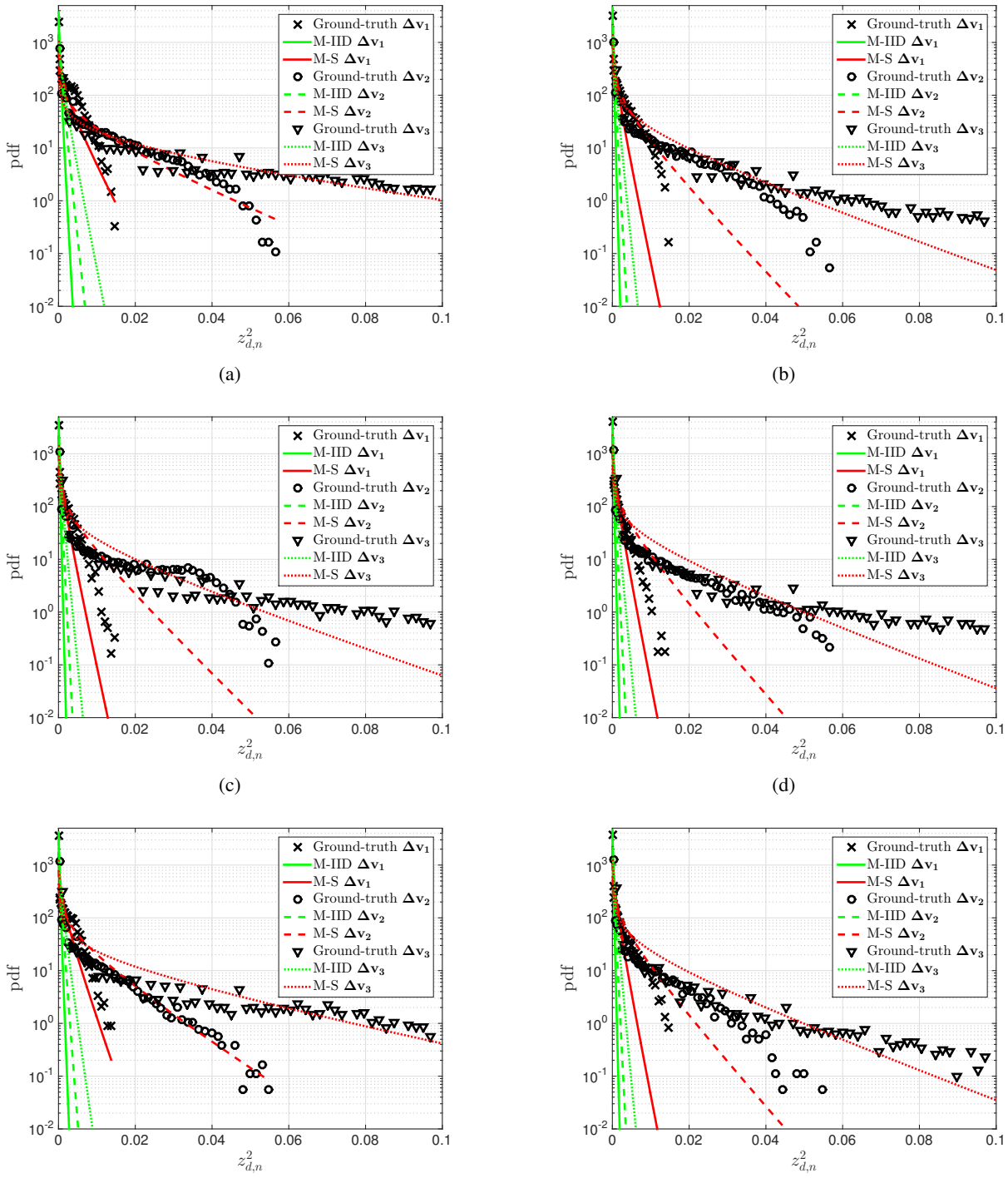

Fig. 5: Estimated and ground-truth distributions for $z^2_{d,n}$, using M-IID and M-S, with translation errors $\boldsymbol{\Delta}\mathbf{v}_1$, $\boldsymbol{\Delta}\mathbf{v}_2$ and $\boldsymbol{\Delta}\mathbf{v}_3$, using the SMVS dataset. In these plots, (a) $d = 1$, (b) $d = 2$, (c) $d = 3$, (d) $d = 4$, (e) $d = 5$, and (f) $d = 6$, all with $n = 3$.

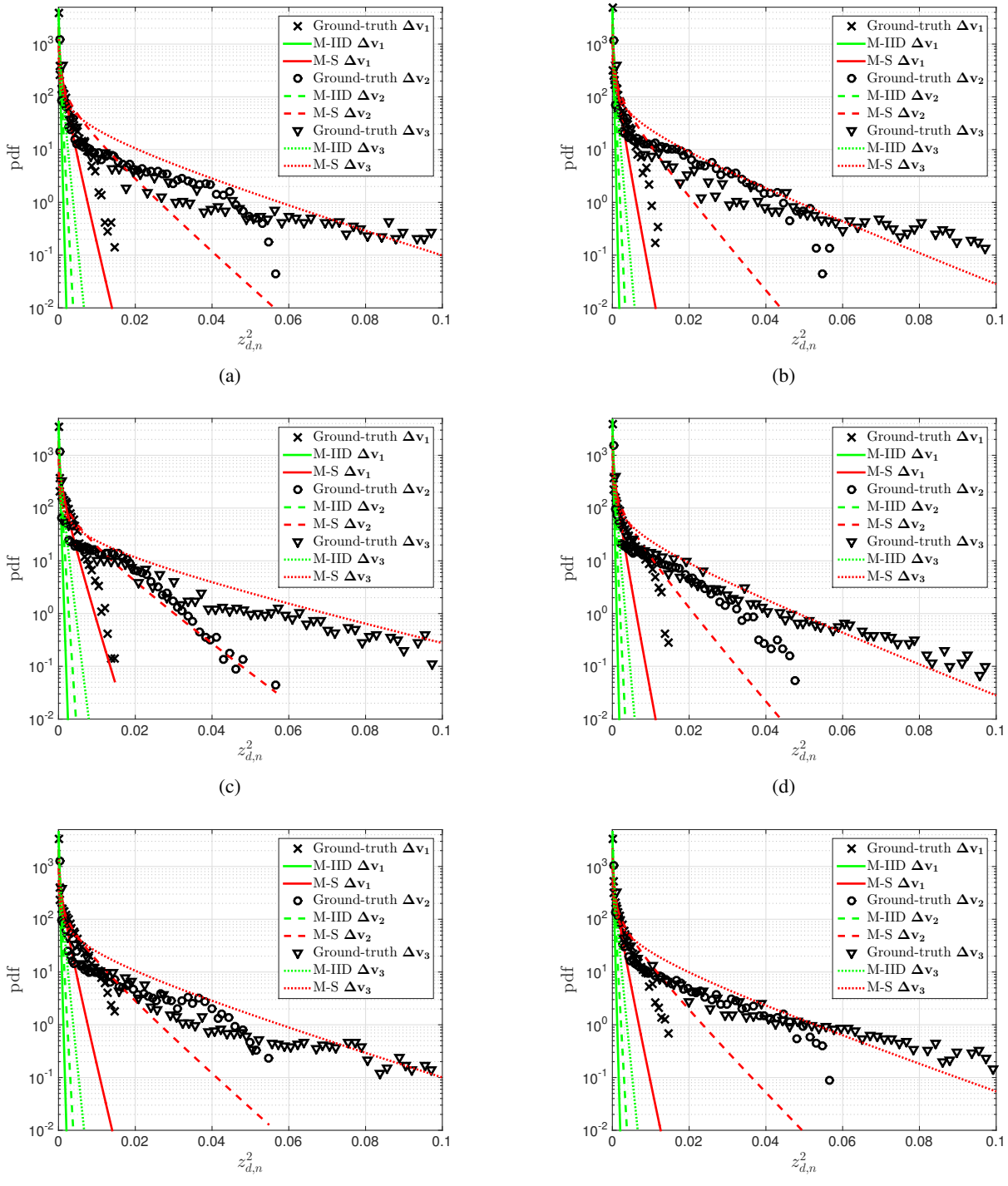

Fig. 6: Estimated and ground-truth distributions for $z^2_{d,n}$, using M-IID and M-S, with translation errors $\mathbf{\Delta v}_1$, $\mathbf{\Delta v}_2$ and $\mathbf{\Delta v}_3$, using the CNN2h dataset. In these plots, (a) $d = 3$, (b) $d = 4$, (c) $d = 5$, (d) $d = 6$, (e) $d = 7$, and (f) $d = 8$, all with $n = 10$.