# DarSwin-Unet: Distortion Aware Architecture-Supplementary Material

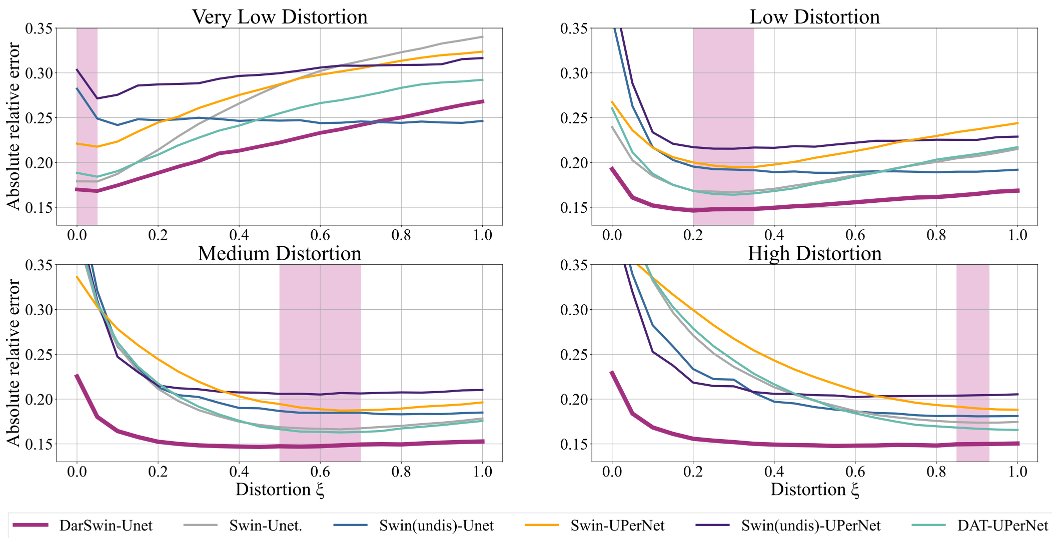


Figure 1. Absolute relative error (lower better) in depth estimation as a function of test distortion for the six baselines: DarSwin-Unet, Swin-Unet [2], Swin(undis)-Unet, Swin-UPerNet [3, 6], Swin(undis)-UPerNet and DAT-UPerNet [5, 6]. All methods are trained on a restricted set of lens distortion curves (indicated by the pink shaded regions): (a) very low, (b) low, (c) medium, and (d) high distortion. We study the generalization abilities of each model by testing across all $\xi \in [0, 1]$. The squared relative error follows the same curves as the absolute relative error.

# A. Depth estimation

## A.1. Evaluation metrics

These detailed explanation of the evaluation metrics as shown in main text are defined as follows:

- Absolute relative error $= \frac{1}{|D|}\sum_{d\in D}\frac{|d^*-d|}{d^*}$
- RMSE $= \sqrt{\frac{1}{|D|}\sum_{d\in D}||d^*-d||^2}$
- Square relative error $= \frac{1}{|D|}\sum_{d\in D}\frac{||d^*-d||^2}{d^*}$
- log-RMSE $= \sqrt{\frac{1}{|D|}\sum_{d\in D}||\log d^*-\log d||^2}$
- $\delta_t = \frac{1}{|D|}|\{d \in D|\max(\frac{d^*}{d}, \frac{d}{d^*}) \leq 1.25^t\}|, \quad t \in \{1, 2, 3\}$

with $D$, $d^*$ and $d$ are respectively the set of valid depths, the ground truth depth and the predicted depth. We show the results for each metric similar to Fig. 9 in the main text.

## A.2. Proposed sampling function

The goal is to identify a class of functions that is parameterized by a minimal number of parameters while still being capable of representing a wide variety of monotonic profiles between two interpolation points, $(0, 0)$ and $(a, b)$. The initial approach involves using a power-law function, which is widely employed in engineering due to its simplicity and its ability to model relationships between unknown quantities with minimal parametrization.

$$p_n(\theta) = b\left(\frac{\theta}{a}\right)^n,$$

The function $p(0) = 0$ and $p(a) =$ FOV. This formulation generates convex curves $(n \geq 1)$ or concave curves $(n < 1)$, with a derivative of zero or a non-existent derivative (tangent to the y-axis) at the origin. The underlying idea is that if a curve exhibits cuspidal behavior at one end, it should also be capable of exhibiting such behavior at the other end. To achieve this symmetry, two reflections are

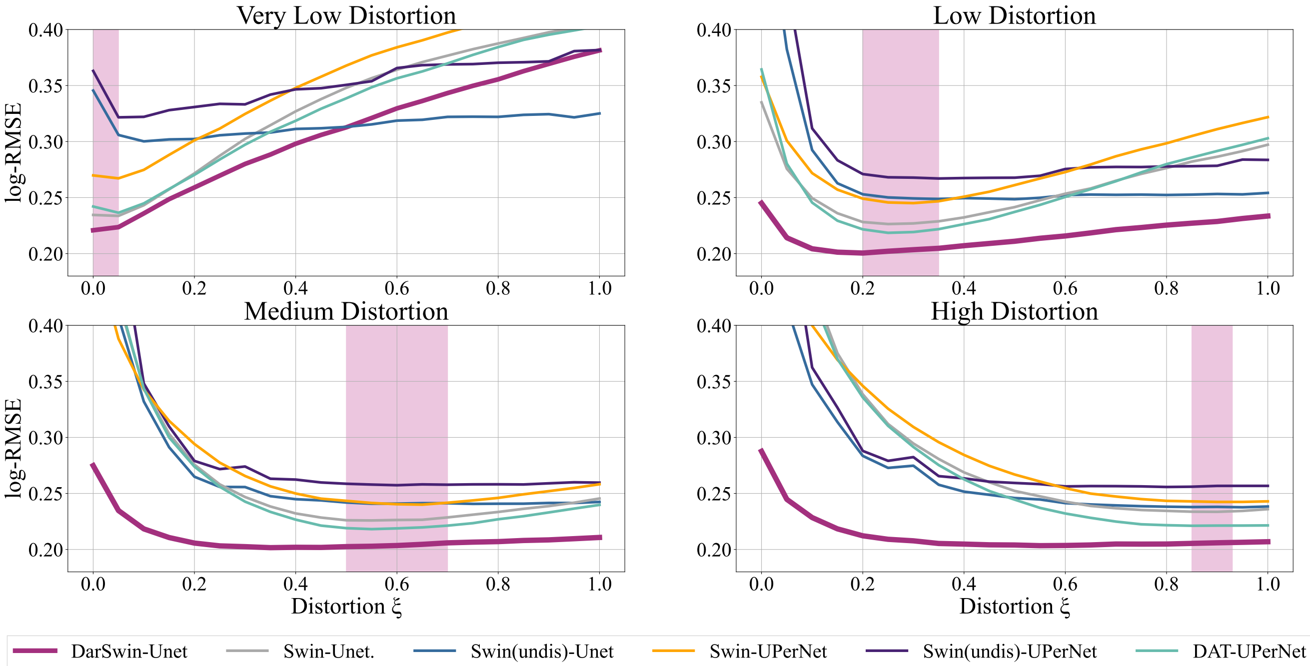


Figure 2. log-RMSE (lower better) in depth estimation as a function of test distortion for the six baselines: DarSwin-Unet, Swin-Unet [2], Swin(undis)-Unet, Swin-UPerNet [3, 6], Swin(undis)-UPerNet and DAT-UPerNet [5, 6] . All methods are trained on a restricted set of lens distortion curves (indicated by the pink shaded regions): (a) Very low, (b) low, (c) medium and (d) high distortion. We study the generalization abilities of each model by testing across all $\xi \in [0, 1]$. RMSE follows the same curves as log-RMSE.

applied to flip the function vertically and horizontally.

$$q_m(\theta) = 1 - \left(1 - \frac{\theta}{a}\right)^m ,$$

which also satisfy the interpolation conditions. This approach can generate both convex ($m \geq 1$) and concave ($m < 1$) curves. To combine these curves while ensuring the interpolation conditions remain satisfied, their convex combination is utilized.

$$g(\theta) = \lambda p_n(\theta) + (1 - \lambda) q_m(\theta) , \tag{1}$$

for $\lambda \in [0, 1]$. If $m < 1$, $n > 1$ or $m > 1$, $n < 1$, the resulting curve is clearly monotonic increasing. In cases where both $m$ and $n$ are either greater than 1 or less than 1, the curve remains monotonic. This family of curves is parameterized by the three parameters $m, n, t$.

## A.3. Derivative of uniform camera model projection

The Unified camera model [1,4] as explained in the main text, *bounded* parameter $\xi \in [0, 1]$[1] projects the world point to the image as follows

$$r_d = \mathcal{P}(\theta) = \frac{f \cos\theta}{\xi + \sin\theta} , \tag{2}$$

[1] $\xi$ can be slightly greater than 1 for certain types of catadioptric cameras [7] but this is ignored here.

where $r_d$ is the radial distance from the image center, $\theta$ the incident angle lens, $f$ the focal length and $\xi$ the distortion parameter.

For a fixed $\theta$ (field of view), to prove that the extremities of the derivatives with respect to $g(\theta)$ for this projection function lie at $\xi = 0$ and $\xi = 1$, we first need to calculate $\frac{\mathrm{d}r_d}{\mathrm{d}g(\theta)}$. To do so, let us first compute $\frac{\mathrm{d}r_d}{\mathrm{d}\theta}$,

$$\frac{\mathrm{d}r_d}{\mathrm{d}\theta} = \frac{\mathrm{d}}{\mathrm{d}\theta}\Big(\frac{f\cos\theta}{\xi + \sin\theta}\Big) , \tag{3}$$

$$\frac{\mathrm{d}r_d}{\mathrm{d}\theta} = \frac{\frac{\mathrm{d}}{\mathrm{d}\theta}(f\cos(\theta))(\xi + \sin(\theta)) - (f\cos(\theta))\frac{\mathrm{d}}{\mathrm{d}\theta}(\xi + \sin(\theta))}{(\xi + \sin(\theta))^2} , \tag{4}$$

$$\frac{\mathrm{d}r_d}{\mathrm{d}\theta} = \frac{-f(\xi\sin\theta + 1)}{(\xi + \sin(\theta))^2} . \tag{5}$$

To calculate $\frac{\mathrm{d}r_d}{\mathrm{d}g(\theta)}$, we can write $\theta = g^{-1}(g(\theta))$ and use chain rule :

$$
\begin{aligned}
\frac{\mathrm{d}\theta}{\mathrm{d}g(\theta)} &= \frac{1}{g'(\theta)}\,, && (6)\\
\frac{\mathrm{d}r_d}{\mathrm{d}g(\theta)} &= \frac{\mathrm{d}r_d}{\mathrm{d}\theta}\frac{\mathrm{d}\theta}{\mathrm{d}g(\theta)}\,, && (7)\\
\frac{\mathrm{d}r_d}{\mathrm{d}g(\theta)} &= \frac{-f(\xi\sin\theta+1)}{g'(\theta)(\xi+\sin(\theta))^2}\,.
\end{aligned}
$$

To prove that the extremities of the derivative of the projection function occur at $\xi = 0$ and $\xi = 1$, we need to prove that the derivative is monotonic, i.e. $\frac{\mathrm{d}}{\mathrm{d}\xi}(\frac{\mathrm{d}r_d}{\mathrm{d}g(\theta)}) > 0$, Fist we analysis this derivative, using the quotient rule, the derivative becomes:

$$
\frac{\mathrm{d}}{\mathrm{d}\xi}\left(\frac{-f(\xi\sin\theta+1)}{g'(\theta)(\xi+\sin\theta)^2}\right) = \frac{-f\sin\theta(\xi+\sin\theta)^2+2f(\xi\sin\theta+1)(\xi+\sin\theta)}{g'(\theta)(\xi+\sin\theta)^4}\,.
$$

The denominator is $g'(\theta)(\xi+\sin\theta)^4$, since $g(\theta)$ is monotonic $g'(\theta) > 0$ and $(\xi+\sin\theta)^4 > 0$.

The numerator is:

$$
\begin{aligned}
N(\xi) = &-f\sin\theta(\xi+\sin\theta)^2 \\
&+2f(\xi\sin\theta+1)(\xi+\sin\theta)\,.
\end{aligned} \tag{8}
$$

Let us analyze this numerator, we want $N(\xi) > 0$ as well,

$$
\begin{aligned}
&2f(\xi\sin\theta+1)(\xi+\sin\theta) > f\sin\theta(\xi+\sin\theta)^2\,, && (9)\\
&2(\xi\sin\theta+1) > \sin\theta(\xi+\sin\theta) \text{ since } ((\xi+\sin\theta)\neq 0)\,,\\
&\xi\sin\theta-\sin^2\theta+2 > 0\,,\\
&2 > \sin\theta(\sin\theta-\xi)\,.
\end{aligned}
$$

Since $\theta = \mathrm{FOV}/2$ it follows that $\theta \in [0,\pi]$, $\sin\theta \in [0,1]$ and $\xi \in [0,1]$. Therefore, the maximum value of $\sin\theta(\sin\theta-\xi)$ occurs at $\xi = 0$ and $\sin\theta = 1$.

Therefore, $\frac{\mathrm{d}}{\mathrm{d}\xi}(\frac{\mathrm{d}r_d}{\mathrm{d}g(\theta)}) > 0$, meaning the derivative $\frac{\mathrm{d}r_d}{\mathrm{d}g(\theta)}$ is monotonic with respect to $\xi$. Consequently, the maximum value of this derivative $\frac{\mathrm{d}r_d}{\mathrm{d}g(\theta)}$ occurs either at $\xi = 0$ or $\xi = 1$.

## References


[1] João P. Barreto. A unifying geometric representation for central projection systems. 2006. 2

[2] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. 2022. 1, 2

[3] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. 2021. 1, 2

[4] Christopher Mei and Patrick Rives. Single view point omnidirectional camera calibration from planar grids. 2007. 2

[5] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. 2022. 1, 2

[6] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. 2018. 1, 2

[7] Xianghua Ying and Zhanyi Hu. Can we consider central catadioptric cameras and fisheye cameras within a unified imaging model. 2004. 2