

EGformer: Equirectangular Geometry-biased Transformer for 360 Depth Estimation

Ilwi Yun¹ Chanyong Shin² Hyunku Lee¹ Hyuk-Jae Lee¹ Chae Eun Rhee²

¹ Seoul National University, Seoul, Republic of Korea

²Inha University, Incheon, Republic of Korea

yuniw@capp.snu.ac.kr, scyongg@inha.edu, hyunku.lee@snu.ac.kr,
 hjlee@capp.snu.ac.kr, chae.rhee@inha.ac.kr

Abstract

Estimating the depths of equirectangular (i.e., 360°) images (EIs) is challenging given the distorted $180^\circ \times 360^\circ$ field-of-view, which is hard to be addressed via convolutional neural network (CNN). Although a transformer with global attention achieves significant improvements over CNN for EI depth estimation task, it is computationally inefficient, which raises the need for transformer with local attention. However, to apply local attention successfully for EIs, a specific strategy, which addresses distorted equirectangular geometry and limited receptive field simultaneously, is required. Prior works have only cared either of them, resulting in unsatisfactory depths occasionally. In this paper, we propose an equirectangular geometry-biased transformer termed EGformer. While limiting the computational cost and the number of network parameters, EGformer enables the extraction of the equirectangular geometry-aware local attention with a large receptive field. To achieve this, we actively utilize the equirectangular geometry as the bias for the local attention instead of struggling to reduce the distortion of EIs. As compared to the most recent EI depth estimation studies, the proposed approach yields the best depth outcomes overall with the lowest computational cost and the fewest parameters, demonstrating the effectiveness of the proposed methods.

1. Introduction

Estimating the depths of equirectangular (i.e., 360°) images (EIs) can be challenging because such images have a $180^\circ \times 360^\circ$ wide field-of-view (FoV) with distortion. Images with distorted wide FoV often requires a global view for proper image processing [20, 26, 53]. Such circumstances strongly require a large receptive field for accurate depth estimations of EIs, which is hard to be achieved via

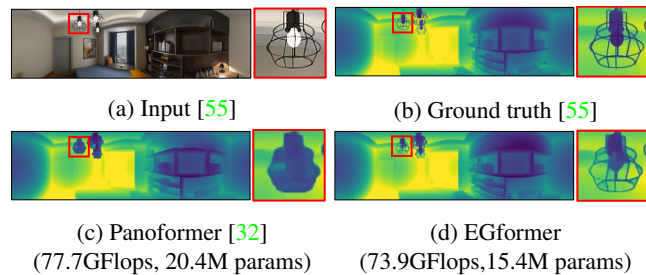


Figure 1: By utilizing the equirectangular geometry as the bias, EGformer efficiently enables the extraction of the equirectangular geometry-aware local attention with a large receptive field, yielding much more accurate depths with the lowest computational cost and the fewest parameters as compared to the results in the most recent studies.

convolutional neural network (CNN).

Considering the receptive field, the vision transformer (ViT) [14] may be the best option for equirectangular depth estimations. ViT has advantages over a CNN in that the attention is extracted in a global manner. A wide FoV can be addressed through global attention, and the effectiveness of this approach has been demonstrated [53]. However, in terms of the computational cost, this global mechanism makes ViT inappropriate for application to EIs. The computational cost of global attention is quadratic with respect to the input resolutions [23] which raises the need for local attention [23, 22, 13, 18, 19].

Unfortunately, applying local attention to EIs is a non-trivial problem because the distorted geometry and local receptive field should be addressed at the same time. Due to the non-uniform geometry of EIs, each local attention should be extracted differently while considering the equirectangular geometry. Therefore, local attention for general vision tasks cannot yield satisfactory performance

outcomes for EIs, which demands specific strategies. However, even if the distortion of EIs can be addressed via proper strategies as proposed recently by Panoformer [32], there still remains a fundamental limitation of local attention: the limited receptive field. To enlarge the receptive field, various hand-crafted or data-adaptive sparse patterns have been proposed for local windows with a hierarchical type of network architecture [42, 43, 23, 13, 45, 32]. However, such repeated local operations cannot fundamentally substitute for a global operator [44], increasing both the computational cost and the number of network parameters required for a plausible quality of the depth.

As described above, dealing efficiently with equirectangular geometry and a limited receptive field via local attention appears to be challenging, yet one important fact has been overlooked: **the equirectangular geometry is known beforehand**. For various vision tasks, it has been shown that a structural prior can boost the performances efficiently [20, 46, 9, 16, 50, 48]. For example, based on the prior knowledge that cars cannot fly up in the sky in urban scenes, HA-Net [9] improves segmentation performance outcomes at a negligible computational overhead by imposing different importance levels on the encoded features according to their vertical positions. Inspired by those studies, we come up with the idea of offsetting the limitations of local attention via a structural prior of EIs.

In this paper, we propose an equirectangular geometry-biased transformer, termed EGformer, which actively utilizes the equirectangular geometry as the bias for inter- and intra-local windows. Through this, while limiting the computational cost and the number of network parameters, EGformer enables the extraction of the equirectangular geometry-aware local attention with a large receptive field. EGformer consists of three main proposals: equirectangular relative position embedding (ERPE), distance-based attention score (DAS) and equirectangular-aware attention rearrangement (EaAR). Specifically, ERPE and DAS impose geometry bias onto the elements within the local window, allowing for consideration of the equirectangular geometry when extracting the local attention. Meanwhile, EaAR imposes the geometry bias on the local window. This enables each local window to interact with other local windows indirectly, thereby enlarging the receptive field. Compared to the most recent studies of EI depth estimations, EGformer yields the best depth outcomes overall with the lowest computational cost and the fewest parameters, demonstrating the effectiveness of the proposed method.

2. Background and related work

2.1. Equirectangular geometry

As shown in Figure 2, EIs are constructed by projecting a sphere image onto two-dimensional (2D) plane, and vice

versa. Therefore, spherical coordinates are used for EIs, and each pixel location is represented through (ρ, θ, ϕ) , where $\theta \in (0, 2\pi)$, $\phi \in (0, \pi)$. Spherical coordinates can be converted to Cartesian coordinates (X, Y, Z) via Eq.(1).

$$\begin{cases} X = \rho \cdot \sin(\phi) \cdot \cos(\theta) \\ Y = \rho \cdot \sin(\phi) \cdot \sin(\theta) \\ Z = \rho \cdot \cos(\phi) \end{cases} \quad (1)$$

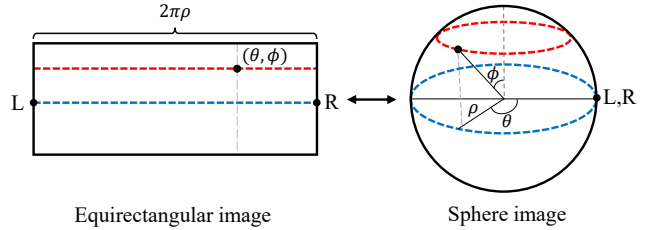


Figure 2: Conversion between EI and sphere image

Structural prior Due to the equirectangular geometry, EIs have some distinct characteristics which should be further considered for proper image processing. Below are several examples. **First**, the information density of EIs differ according to the locations. The information density is low around near $\phi = 0, \pi$, while it is high at $\phi = \frac{\pi}{2}$. The red and blue lines in Figure 2 represent the differences in the information density. Despite having the same number of pixels, the red and blue lines in EIs contain different amount of information, as shown in the sphere image of Figure 2. Therefore, even with equal local window sizes, there exist differences in information quantity according to the locations of local window. **Second**, EIs are cyclic. In other words, the left and right ends of EIs are actually connected although they appear to be separated in EIs. The L and R points in Figure 2 visualize this characteristic. As the worst case, a single object is often split into left and right ends of EIs, requiring specific strategy [25].

2.2. Transformer for vision tasks

Compared to a CNN, ViT [14] possesses a global receptive field, which is highly beneficial for various vision tasks [6, 17, 52, 47, 28, 21]. However, due to the high computational cost of global attention, several studies have focused on utilizing the local attention based on hierarchical architecture [42, 43]. The Swin Transformer (SwinT) [23] proposes square-shaped local attention and the associated shifting mechanism, and Deformable attention transformer [45] further improves SwinT through deformable attention inspired by the deformable convolution [11, 56]. However, square-shaped local attention with a hierarchical architecture enlarges the receptive field too slowly. To alleviate this,

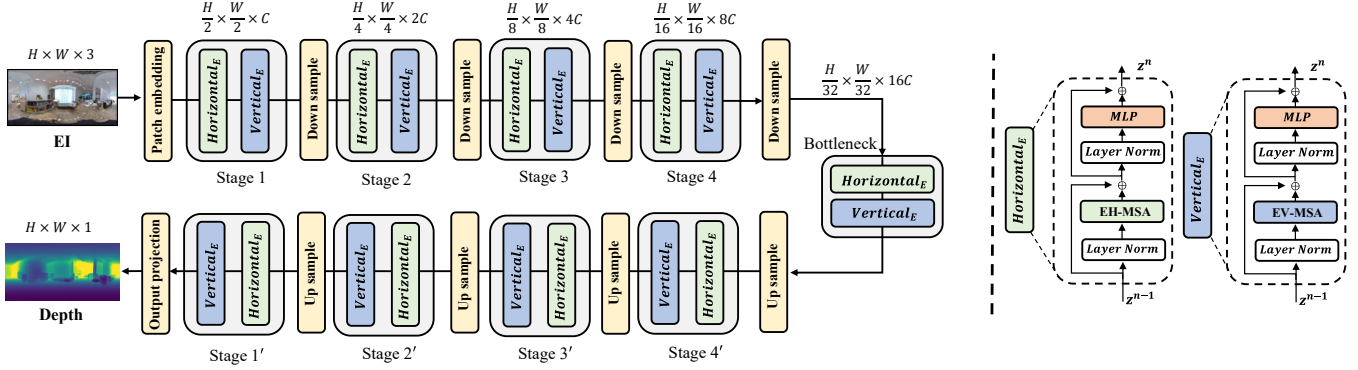


Figure 3: The network architecture of EGformer variants (left) and EGformer transformer block (right). Stage n and Stage n' indicates encoder and decoder, respectively. For better visualization, skip connection[31] between encoder and decoder is omitted here, and more details are included in **Technical Appendix**. $Horizontal_E$ and $Vertical_E$ represents the proposed transformer block of EGformer (Section 3.2), which comprise equirectangular-aware vertical and horizontal multi-head self-attention (*i.e.*, EH-MSA, EV-MSA) (Section 3.3).

the CSwin transformer (CSwinT) [13] proposes the use of horizontal and vertical local attention in parallel, and Dilateformer [19] proposes dilated attention inspired by dilated convolution [51]. Nevertheless, both remain limited in that the receptive field is bounded to the sizes of the local windows.

2.3. 360 monocular depth estimation

The topics of equirectangular depth estimation studies mostly fall into two categories: Dealing with equirectangular geometry or dealing with insufficient data. To address the distortion of EIs, several studies [39, 40] utilize cubemap projections with certain padding schemes [8]. Convolution kernels considering equirectangular geometry [10, 33] have also been studied. Instead of addressing the distortion directly, some studies utilize the equirectangular geometry to improve the performance. Based on the finding that the geometric structures of EIs are embedded along the vertical direction [12], Hohonet [35] and SliceNet [26] propose to process EIs in a vertical direction. Jin *et al.* [20] and Zeng *et al.* [54] show that some prior knowledge of geometric structure of EIs can boost the performances further. Meanwhile, due to the distorted and wide FoV, acquiring ground truth equirectangular depths is extremely difficult, resulting in lack of data [24, 58, 53]. Therefore, some studies have attempted to address data insufficiency through self-supervised learning [38, 57, 25, 41] or transfer learning [53]. Recently, inspired by the success of ViT [14], there have been several attempts to apply a transformer to equirectangular depth estimations. Yun *et al.* [53] demonstrated that global attention can effectively handle the wide FoV of EIs. However, global attention is computationally inefficient and requires pre-training on a large-scale dataset to perform at its best. To address this issue, Panoformer

[32] proposed pixel-based local attention for which the calculations are done by sampling nearby pixels according to the equirectangular geometry. To manage the small receptive field of local attention, Panoformer adaptively adjusts local window sizes via a learnable offset, similar to that of a deformable mechanism [11]. However, because training accurate and large offsets for a deformable mechanism is extremely difficult in practice [56, 45], Panoformer is also associated with a limited receptive field.

3. EGformer

3.1. Overview

Figure 3 illustrates the architecture of an EGformer variant (refer to Section 4.2 for more variants). Each $Horizontal_E$ (green block) and $Vertical_E$ (blue block) represents the proposed horizontal and vertical transformer blocks of EGformer (Section 3.2), which comprises equirectangular-aware horizontal and vertical multi-head self-attention (EH-MSA, EV-MSA), as illustrated on the right side of Figure 3 (Section 3.3). The yellow blocks (*e.g.*, Patch embedding, Down sample) in Figure 3 are based on a CNN. Details are included in the **Technical Appendix**.

3.2. Horizontal and vertical transformer block of EGformer

As discussed in Section 2.3, EIs have distinct natures along the vertical and horizontal directions [12]. The geometric structure (*e.g.*, layout) is embedded along the vertical direction [53, 49, 34], while the cyclic structure of EIs can be addressed implicitly along the horizontal direction. For these reasons, prior studies on EIs have leveraged these natures to enhance their performance [35, 26]. Drawing insights from these work, we adopt vertical and horizontal

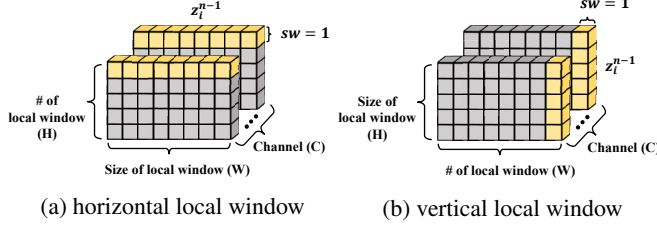


Figure 4: Local window shapes of EGformer

shaped local window for EGformer.

Let us define $z^{n-1} \in \mathbb{R}^{H \times W \times (d_j \times J)}$ as the output of $(n-1)$ -th transformer block or precedent convolutional layer, which are input to the n -th transformer block. Here, H and W is the height and width, where J represents the number of heads in multi-head self-attention (MSA) and d_j indicates the number of hidden layers of each head. In total, the channel dimension of z^{n-1} is calculated via $C = d_j \times J$.

When z^{n-1} inputs to the $Horizontal_E$, z^{n-1} is divided along the horizontal direction with stripe width (sw) as 1 (Figure 4 (a)), constructing the group of horizontal local window features as formulated by $[z_1^{n-1}, z_2^{n-1}, \dots, z_H^{n-1} \in \mathbb{R}^{1 \times W \times C}]$ in Eq.(2). Through layer normalization (LN), the normalized features of i -th horizontal local window for j -th head (i.e., $F_i^j \in \mathbb{R}^{1 \times W \times d_j}$) is extracted via Eq.(3). Then, query, key, value (i.e., $Q_i^j, K_i^j, V_i^j \in \mathbb{R}^{1 \times W \times d_j}$) are obtained by linearly projecting the F_i^j as described in Eq.(4). Afterwards, through proposed EH-MSA, the local attention for i -th horizontal local window $L_i^j \in \mathbb{R}^{1 \times W \times d_j}$ is extracted. By accumulating L_i^j along the height and head dimension, equirectangular-aware horizontal attention $L \in \mathbb{R}^{H \times W \times C}$ is constructed as shown in Eq.(5). Finally, following the previous works [23, 13, 45], the output of n -th $Horizontal_E$ (z_n) is defined by Eq.(6).

$$[z_1^{n-1}, z_2^{n-1}, \dots, z_H^{n-1}] = z^{n-1} \quad (2)$$

$$[F_i^j]^{j=1, \dots, J} = LN(z_i^{n-1}) \quad (3)$$

$$Q_i^j, K_i^j, V_i^j = Linear(F_i^j) \quad (4)$$

$$\begin{aligned} L_i^j &= EH\text{-MSA}(Q_i^j, K_i^j, V_i^j) \\ L &= [L_1^j, L_2^j, \dots, L_H^j]^{j=1, \dots, J} \end{aligned} \quad (5)$$

$$\begin{aligned} \hat{z}^n &= L + z^{n-1} \\ z^n &= MLP(LN(\hat{z}^n)) + \hat{z}^n \end{aligned} \quad (6)$$

In the same vein, the final output of n -th $Vertical_E$ is extracted equally through EV-MSA with the only difference being that the group of vertical local window features (i.e.,

$[z_1^{n-1}, z_2^{n-1}, \dots, z_W^{n-1} \in \mathbb{R}^{1 \times H \times C}]$) is made by dividing z^{n-1} along vertical direction (Figure 4 (b)).

3.3. Equirectangular-aware horizontal and vertical self-attention

The overall process of EH-MSA is illustrated in Figure 5. When calculating the attention score ($QK^T + E(\phi)$), ERPE ($E(\phi)$) is added to QK^T similar to relative position embedding [23]. Then, the Das (blue block) is calculated from attention score, which produces the attention for the current block. Additionally, the importance level of each local window (M^h) is obtained from the attention score. Finally, through EaAR, the final attention of EH-MSA (L) is provided. The following subsections describe each component of E(V)H-MSA in detail.

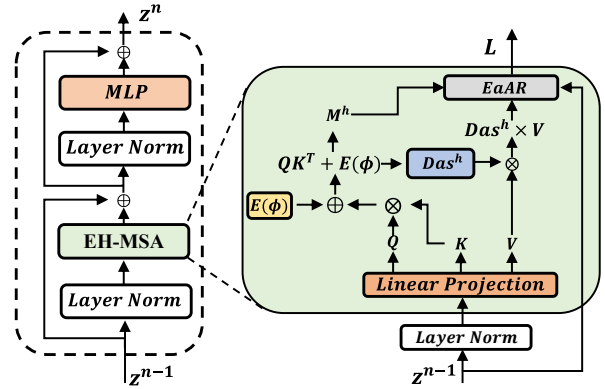


Figure 5: Overall process of EH-MSA. The components in yellow (ERPE), blue (Das) and gray block (EaAR) denote our main proposals.

Equirectangular relative position embedding We propose a non-parameterized ERPE to impose equirectangular geometry bias on the elements within each vertical and horizontal local windows. We define $E(\phi_i) \in \mathbb{R}^{W \times W}$ as the ERPE for horizontal local windows, where the (m, n) -th element of $E(\phi_i)$ is expressed via $E(\phi_i)_{m,n} \in \mathbb{R}^{1 \times 1}$. The calculation process of $E(\phi_i)_{m,n}$ is defined by Eq.(7). Here, $\theta_{m,n}$ denotes the positions of the m -th and n -th elements in the horizontal local windows, where ϕ_i denotes the positions of the i -th horizontal local windows. Similarly, the ERPE for the i -th vertical local window is expressed via $E(\theta_i) \in \mathbb{R}^{H \times H}$, where the (m, n) -th element of $E(\theta_i)_{m,n} \in \mathbb{R}^{1 \times 1}$ is calculated with Eq.(8)¹. The $\text{sign}(\cdot)$ function is used to distinguish between $E(\phi_i)_{m,n}$ and $E(\phi_i)_{n,m}$.

$$E(\phi_i)_{m,n} = \text{sign}(\theta_m - \theta_n) \cdot \rho \sqrt{2\{1 - \cos(\theta_m - \theta_n)\}} \cdot \sin(\phi_i) \quad (7)$$

¹In experiments, we set $\rho = 0.1$. Refer to Table 5 for more details.

$$E(\theta_i)_{m,n} = \text{sign}(\phi_m - \phi_n) \cdot \rho \sqrt{2\{1 - \cos(\phi_m - \phi_n)\}} \quad (8)$$

ERPE is calculated by measuring the distances between the m -th and n -th elements in Spherical coordinates as illustrated in Figure 6. The green line in Figure 6 denotes $E(\phi_i)$ of each local window, while each blue and red line represents the corresponding horizontal local window. Unlike position embedding for general vision tasks, ERPE can enforce the attention score to assign high similarity for the elements that are close in three-dimensional space. For instance, the ERPE of L and R in Figure 2 is 0 although they are far apart in EIs. As a result, ERPE induces the transformer to assign similar attention score for L and R that makes transformer to understand the cyclic structure of EIs.

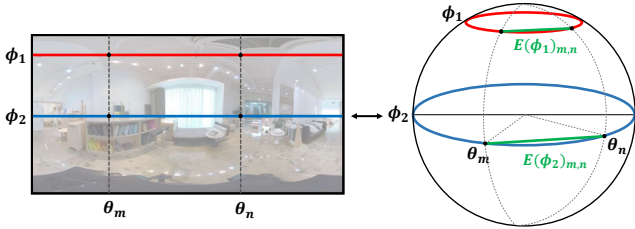


Figure 6: ERPE for horizontal local window. Each red and blue line represents the horizontal local window. Green line indicates the ERPE which are calculated through the distance in Cartesian coordinate.

Distance-based attention score Conventionally, softmax has been preferred for re-weighting the attention score. However, softmax is computationally inefficient [3, 30, 4, 36, 27] and lacks relevance to the natural language processing (NLP) or vision tasks. For these reasons, some studies have attempted to replace softmax with alternative functions [36, 27]. These studies empirically determined that the essential roles performed by softmax are as follows.

- Ensuring attention score to get non-negative value
- Re-weighting the attention score

Inspired by their findings, we propose a distance-based attention score for each i -th horizontal ($Das_i^h \in \mathbb{R}^{1 \times W \times W}$) and vertical ($Das_i^v \in \mathbb{R}^{1 \times H \times H}$) local window, defined by Eqs.(10) and (11) respectively. Here, \mathcal{N} represents L_1 normalization, and $(\rho_b, \theta_b, \phi_b)$ represents the baseline point (hyperparameter). Then, the local attention of each i -th local window is obtained using Eq.(12)

$$\begin{aligned} score_i^h &= Q_i K_i^T + E(\phi_i) \\ score_i^v &= Q_i K_i^T + E(\theta_i) \end{aligned} \quad (9)$$

$$Das_i^h = 2\rho_b^2 \cdot \{1 - \cos(\mathcal{N}\{score_i^h\} \cdot \frac{\pi}{2})\} \cdot \sin^2(\phi_b) \quad (10)$$

$$Das_i^v = 2\rho_b^2 \cdot \{1 - \cos(\mathcal{N}\{score_i^v\} \cdot \frac{\pi}{2})\} \quad (11)$$

$$\begin{aligned} Attention_i^h &= Das_i^h \times V_i \\ Attention_i^v &= Das_i^v \times V_i \end{aligned} \quad (12)$$

The core idea of $Das_i^{h,v}$ is to convert each element of $score_i^{h,v}$ into the distances from the baseline point $(\rho_b, \theta_b, \phi_b)$ in Spherical coordinates. Simply put, the farther the element of $score_i^{h,v}$ is from $(\rho_b, \theta_b, \phi_b)$, the higher the distance-based attention score it receives. In this paper, we set baseline point as $(\rho_b, \theta_b, \phi_b) = (\frac{1}{\sqrt{2}}, 0, \frac{\pi}{2})$ to make both Das^h and Das^v get equal score range $[0, 1]$. The calculation process of Das_i^h is visualized in Figure 7, which is performed via the following steps. First, through normalization as denoted by the black arrow in Figure 7, $score_i^h$ is converted to $\Delta\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$, as visualized by the green curve in Figure 7. Second, by calculating the distance of $(\frac{1}{\sqrt{2}}, 0 + \Delta\theta, \frac{\pi}{2})$ from $(\frac{1}{\sqrt{2}}, 0, \frac{\pi}{2})$ in Spherical coordinates, $\sqrt{Das_i^h}$ is obtained, as represented via the purple line in Figure 7. Finally, as shown in Eq.(10), square of the distance is calculated for the final distance-based attention score to focus more on the important region by amplifying the differences in score value. In the same vein, $score_i^v$ is converted to $\Delta\phi \in (-\frac{\pi}{2}, \frac{\pi}{2})$. Then, by calculating the square of the distance of $(\frac{1}{\sqrt{2}}, 0, \frac{\pi}{2} + \Delta\phi)$ from $(\frac{1}{\sqrt{2}}, 0, \frac{\pi}{2})$ in Spherical coordinates, Das_i^v is obtained as described in Eq.(11).

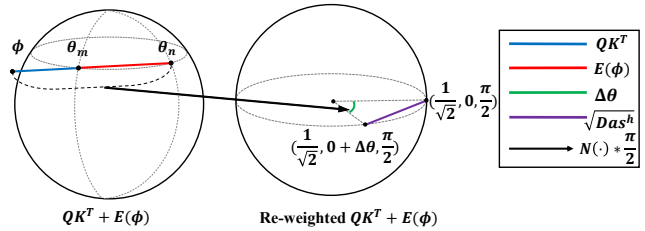


Figure 7: The illustration of the distance-based attention score for horizontal local window. Each element in $score_i^h$ is re-weighted to the distances from the $(\frac{1}{\sqrt{2}}, 0, \frac{\pi}{2})$, which determines the distance-based attention score of each element.

Compared to softmax which is biased in positive values², Das is more appropriate for ERPE. Because $E(\theta, \phi)_{m,n} = -E(\theta, \phi)_{n,m}$, softmax forces unbalanced score for (m, n) -th and (n, m) -th elements when

²To get high softmax attention score, an element in $score$ should have higher 'positive' values than others.

ERPE is imposed. Unlike the NLP in which the order of each element matters, it is often not for depth estimation tasks. Therefore, unbalanced *score* may confuse a transformer, resulting in incorrect score extraction. On the contrary, *Das* is symmetric³. Therefore, when ERPE is used with *Das*, a transformer can get balanced *score* if required, which would be more suitable for EI depth estimation.

Equirectangular-aware attention rearrangement Although structural prior of EIs is embedded in $Attention_i$ via ERPE and *Das*, structural prior between $Attention_i$ is not yet imposed. To address this issue, we propose equirectangular-aware attention rearrangement, defined by Eq.(13). $L_i^h \in \mathbb{R}^{1 \times W \times C}$ and $L_i^v \in \mathbb{R}^{1 \times H \times C}$ indicate the rearranged local attention of the i -th horizontal and vertical local window respectively, with $M_i^h \in \mathbb{R}^{1 \times 1 \times 1}$ and $M_i^v \in \mathbb{R}^{1 \times 1 \times 1}$ representing the importance level of each local window. The importance level of each local window is defined by Eq.(15), where $f(x)$ calculates the mean of all elements in x . Local windows that are important receive M values close to 1, while those that are unimportant receive M values close to 0⁴.

$$\begin{aligned} L_i^h &= M_i^h \cdot (Attention_i) + (1 - M_i^h) \cdot z_i^{n-1} \\ L_i^v &= M_i^v \cdot (Attention_i) + (1 - M_i^v) \cdot z_i^{n-1} \end{aligned} \quad (13)$$

$$f(x) = \frac{\sum_{p=1}^N x_p}{N} \quad (14)$$

$$M_i^h = \frac{f(|score_i^h|)}{\max_{\forall i}(f(|score_i^h|))}, M_i^v = \frac{f(|score_i^v|)}{\max_{\forall i}(f(|score_i^v|))} \quad (15)$$

To approximate $M_i^{h,v}$, we utilize $score_i^{h,v}$. Because $score_i^{h,v}$ is equirectangular geometry-biased, the mean of $|score_i^{h,v}|$ implicitly reflects both the information density and distinctive features of each local window. Specifically, $E(\theta, \phi)$ term in $score_i^{h,v}$ is closely related to the information density⁵ and QK^T term is related to the distinctive characteristics of each local window. Because the information density of local window have high relevance with the level of importance, geometry-biased score values (i.e., $|score_i^{h,v}|$) are appropriate to estimate $M^{h,v}$.

As shown in Eq.(15), the final importance level of i -th local window is obtained by comparing the $f(|score_i^{h,v}|)$ of all other local windows. Therefore, $M_i^{v,h} * Attention_i^{h,v}$

³ $1-\cos(x)=1-\cos(-x)$

⁴ In experiments, we clamp M_i to have value of 0.5 at its minimum to ensure certain amount of $Attention_i$ to be used for L

⁵ $f(|E(\theta_i, \phi_i)|)$ gets proportional relationship with the information density of each i -th local window as shown in Figure 6.

term achieve global-like characteristics by making the local attention to interact with each other indirectly. As a result, the local attention can be extracted more accurately from a feature map with a high resolution. This enables the retention of detailed spatial information and ultimately improves the depth quality. However, in practice, the importance level of each local window can be predicted incorrectly. This could potentially dilute the important information of $Attention_i^{h,v}$ via multiplication with $M_i^{v,h}$. The $(1 - M_i^{v,h}) * z_i^{n-1}$ term can prevent such a situation and enable L_i to be extracted by observing various attention blocks simultaneously, resulting in a more global representation of L_i .

Computational complexity The computational complexity of EH(V)-MSA is as follows:

$$\begin{aligned} \Omega(\text{EH-MSA}) &= 4HW C^2 + 2HW^2 C \\ \Omega(\text{EV-MSA}) &= 4HW C^2 + 2H^2 W C \end{aligned} \quad (16)$$

4. Experiments

Due to page limitation, detailed experimental environment is described in the Technical Appendix.

4.1. Experimental environment

Dataset We evaluate our method using Structured3D [55] and Pano3D [1] datasets, which are the most recent datasets with the highest quality. Discussions on other datasets [5, 2, 58, 38] are included in Technical Appendix.

Metrics The scale of the depth differs according to how the depth is acquired; therefore, an alignment process is commonly used when evaluating the depths of multiple dataset simultaneously [15, 7, 37, 29, 28, 53]. Following earlier work [28], we align the depths in an image-wise manner before measuring the errors for each dataset as defined by Eq.(17) for all methods. Quantitative results are extracted by comparing aligned depth ($Depth_A$) with the ground truth (GT).

$$\begin{aligned} s, t &= \underset{s, t}{\operatorname{argmin}}(s \cdot Depth + t - GT) \\ Depth_A &= s \cdot Depth + t \end{aligned} \quad (17)$$

Common evaluation metrics are used. Lower is better for the absolute relative error (Abs.rel), squared relative error (Sq.rel), root mean square linear error (RMS.lin), root mean square log error (RMSlog). Meanwhile, higher is better for relative accuracy (δ^n), where δ^n represents $\delta < 1.25^n$. FLOPs are calculated using Structured3D testset [55].

ID	Encoder	Decoder	Abs.rel	Sq.rel	RMS.lin	RMSlog	δ^1	δ^2	δ^3	#Param	FLOPs
0	HHHH	HHHH	0.0421	0.0373	0.2961	0.1007	0.9784	0.9922	0.9960	15.3M	81.5G
1	VVVV	VVVV	0.0389	0.0346	0.2983	0.0998	0.9782	0.9920	0.9959	15.3M	70.1G
2	EEEE	EEEE	0.0375	0.0320	0.2945	0.0979	0.9782	0.9920	0.9960	15.3M	75.8G
3	MMMM	MMMM	0.0366	0.0308	0.2795	0.0959	0.9798	0.9926	0.9963	17.5M	80.1G
4	PPEE	EEPP	0.0362	0.0318	0.2874	0.0979	0.9791	0.9921	0.9960	15.6M	77.6G
5	MMEE	EEMM	0.0342	0.0279	0.2756	0.0932	0.9810	0.9928	0.9964	15.4M	73.9G

Table 1: Depth estimation results on network variants for Structured3D testset [55]. Bottleneck layer is fixed to E.

Testset	Method	Backbone	Abs.rel	Sq.rel	RMS.lin	RMSlog	δ^1	δ^2	δ^3	#Param	FLOPs
Structured3D [55]	Bifuse [39]	CNN	0.0644	0.0565	0.4099	0.1194	0.9673	0.9892	0.9948	253.0M	723.4G
	SliceNet [26]	CNN+RNN	0.1103	0.1273	0.6164	0.1811	0.9012	0.9705	0.9867	79.5M	84.3G
	Yun <i>et al.</i> [53]	Global	0.0505	0.0499	0.3475	0.1150	0.9700	0.9896	0.9947	123.7M	589.4G
	Panoformer [32]	Local	0.0394	0.0346	0.2960	0.1004	0.9781	0.9918	0.9958	20.4M	77.7G
	EGformer	Local	0.0342	0.0279	0.2756	0.0932	0.9810	0.9928	0.9964	15.4M	73.9G
Pano3D [1]	Bifuse [39]	CNN	0.1704	0.1528	0.7272	0.2466	0.7680	0.9251	0.9731	253.0M	723.4G
	SliceNet [26]	CNN+RNN	0.1254	0.1035	0.5761	0.1898	0.8575	0.9640	0.9867	79.5M	84.3G
	Yun <i>et al.</i> [53]	Global	0.0907	0.0658	0.4701	0.1502	0.9131	0.9792	0.9924	123.7M	589.4G
	Panoformer [32]	Local	0.0699	0.0494	0.4046	0.1282	0.9436	0.9847	0.9939	20.4M	77.7G
	EGformer	Local	0.0660	0.0428	0.3874	0.1194	0.9503	0.9877	0.9952	15.4M	73.9G

Table 2: Quantitative depth results of each method. Numbers in **bold** indicate the best results. It is observed that EGformer achieves the best depth outcomes with the lowest FLOPs and the fewest parameters.

4.2. Model study

Because all methods have their pros and cons, it is often observed that combining several methods yields better outcomes. For this reason, we study various EGformer variants. Figure 8 shows the various attention module as denoted by **E**, **M**, **P**, **H** and **V**. Here, *PST* indicates Panoformer attention block [32]. Based on this annotation, network architecture can be expressed simply. For example, network architecture in Figure 3 is expressed via 'EEEE-E-EEEE'.

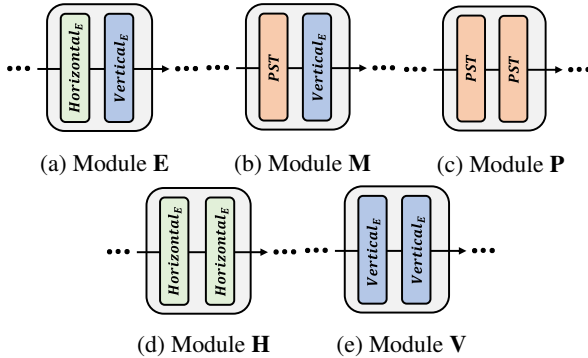


Figure 8: Various attention module used in Table 1. *PST* represents the attention block of Panoformer [32].

Table 1 shows the quantitative results of each network variant for the Structured3D testset [55]. As shown in Table 1, the **H** or **V** attention module does not yield plausible

depth results as shown in ID 0 and 1. As discussed in previous studies [13], the relatively poor performances of **H/V** modules can be explained via their narrow stripe widths (*i.e.*, $sw = 1$). Although consecutive horizontal and vertical attention module (**E**) can alleviate the problem of a narrow stripe width, as shown in ID 2, this solution falls short. Under the circumstances, the easiest means of improving depth quality level is to enlarge the stripe width [13]. However, a wider stripe width also increases the computational cost significantly. Therefore, instead of using wider stripes, we attempt to improve the depth quality by mixing various attention modules, as shown in ID 3, 4 and 5. Among these, we observe that ID 5 is the best fit for our purpose. Based on these results, we set the network architecture of ID 5 as the default architecture in this paper. Meanwhile, the performance differences between ID 3,4 and ID 5 clearly demonstrate the effect of the proposed EH(V)-MSA.

4.3. Comparison with state-of-the-arts

To demonstrate the effectiveness of our proposals, we compare EGformer with the state-of-the-arts. Table 2 shows the quantitative results of each method. Compared to CNN or RNN based approaches, it is observed that transformer-based approaches yield much better depth outcomes. Among which, EGformer achieves the best depth outcomes with the lowest computational cost and the fewest parameters. The reason for low depth qualities of Yun *et al.* [53] could be the lack of dataset. Vision transformer, in which Yun *et al.* based on, requires large-scale dataset

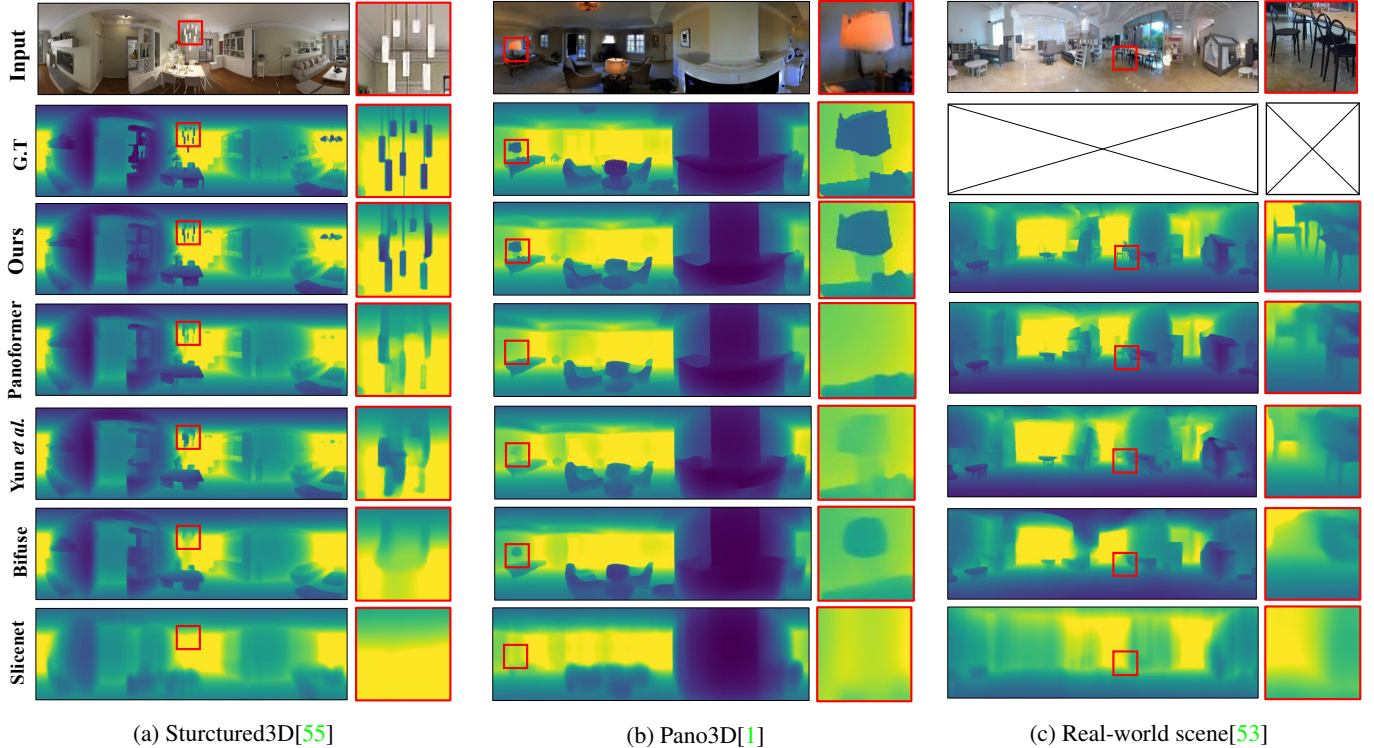


Figure 9: Qualitative results of each method. G.T represents ground truth. Due to page limitations, top-down parts of images are cropped in this figure. There are no ground truth depths in (c) Real-world scene [53].

[14]. However, under our experimental environment, scarce equirectangular depth dataset is used only, which may impair the performances. A further analysis is included in Technical Appendix.

Figure 9 shows the qualitative results of each method. As similar to Table 2, transformer-based approaches provide much better results than that of CNN or RNN based approaches. Slicenet [26] lacks details as reported similarly by [32], and Bifuse [39] provides unsatisfactory results considering the large computational cost and parameters. Out of transformer-based approaches, EGformer yields the best performance in terms of the details. In particular, Panoformer fails to reconstruct the depth of a lamp in Pano3D testset. On the contrary, EGformer reconstructs the depth of a lamp successfully. This result supports our argument in that EGformer extracts the attention accurately even from a feature map with a high resolution, which enables to keep the detailed spatial information. The results on challenging real-world scenes further demonstrate our arguments. All methods fail to distinguish chairs from the background except the EGformer.

4.4. Ablation study

As discussed in Section 3.3, each component of EH(V)-MSA is engineered to perform at its best when they are used

Data	ID	ERPE	Das	EaAR	Abs.rel	Sq.rel	RMSlin	RMSlog	δ^1
S3D	0	✓	✓	✓	0.0342	0.0279	0.2756	0.0932	0.9810
	1	✓		✓	0.0360	0.0307	0.2804	0.0948	0.9799
	2	✓	✓		0.0363	0.0301	0.2804	0.0966	0.9805
	3	✓			0.0374	0.0318	0.2914	0.0984	0.9791
	4				0.0371	0.0316	0.2859	0.0971	0.9793
Pano3D	0	✓	✓	✓	0.0660	0.0428	0.3874	0.1194	0.9503
	1	✓		✓	0.0677	0.0443	0.3972	0.1225	0.9479
	2	✓	✓		0.0687	0.0449	0.3966	0.1228	0.9479
	3	✓			0.0689	0.0448	0.3983	0.1227	0.9482
	4				0.0700	0.0466	0.4052	0.1248	0.9455

Table 3: Ablation study. ID 4 uses softmax attention score with locally enhanced position encoding (LePE) [13]. S3D represents the Structured3D testset [55].

altogether. For example, ERPE requires symmetric characteristics of *Das* to impose geometry bias naturally, and EaAR requires a well-biased *score* to rearrange the attention properly. Table 3 shows these characteristics. Here, ID 4 uses softmax attention score with locally enhanced position encoding (LePE) [13]. Compared to ID 0, a significant performance drop is observed when each component of EH(V)-MSA is removed from ID 0 as shown in ID 1,2,3 and 4. Although these dependencies can be seen as a weakness of EH(V)-MSA, they also suggest that EH(V)-MSA is elaborately designed, which explains why EGformer enables the efficient extraction of the attention for EIs.

Further study on the effect of EH(V)-MSA Because the network architecture of EGformer in Table 3 is 'MMEE-E-EEMM', PST in \mathbf{M} may dilute the effect of EH(V)-MSA. Therefore, to see the effect of EH(V)-MSA more clearly, we conduct an additional ablation study. Table 4 and Figure 10 show the depth estimation results of 'EEEE-E-EEEE' architecture for Structured3D testset. As equal to ID 4 in Table 3, 'Baseline' represents the model that uses softmax attention score with LePE [13] as similar to that of CSwin transformer [13]. As shown in Table 4, improvements are observed when EH(V)-MSA is used instead of CSwin attention mechanism. Meanwhile, Figure 10 shows interesting results. As similar to the result of Panoformer in Figure 9 (b), it is shown that Baseline model fails to reconstruct the depth of a small chair as shown in Figure 10. These results further support our arguments in that the lack of details in depths are common limitation of small receptive field regardless of the shape of the local window. On the contrary, EH(V)-MSA reconstructs the depth of a chair appropriately. This demonstrates clearly in that EH(V)-MSA acts as a key role in keeping the detailed spatial information.

Method	Abs.rel	Sq.rel	RMS.lin	RMSlog	δ^1	δ^2	δ^3
Baseline	0.0399	0.0358	0.3016	0.1014	0.9766	0.9916	0.9958
EH(V)-MSA	0.0375	0.0320	0.2945	0.0979	0.9782	0.9920	0.9960

Table 4: Depth estimation results when different attention mechanism is used for 'EEEE-E-EEEE' architecture for Structured3D testset [55]. Baseline uses softmax attention score with LePE [13].

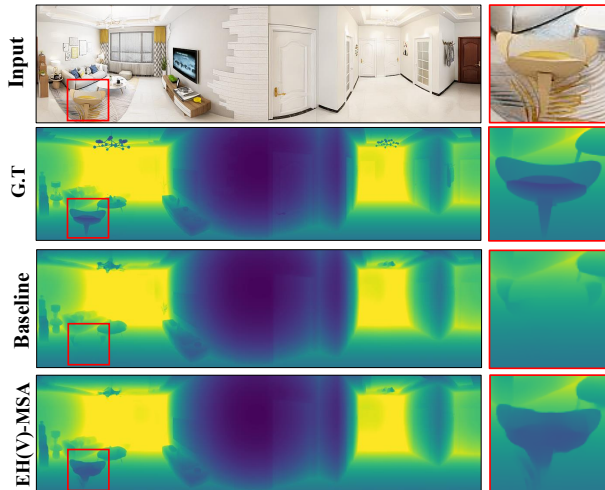


Figure 10: Effect of EH(V)-MSA for 'EEEE-E-EEEE' network architecture. Baseline uses softmax attention score with LePE [13].

Ablation study on bias level (ρ) Although it is demonstrated that equirectangular geometry bias is effective in extracting accurate local attention, excessive geometry bias also may dilute the important information in attention. In EGformer, the influence of equirectangular geometry bias on attention is controlled by ρ in Eqs.(7) and (8). Therefore, to find the appropriate bias level (ρ), we conduct an experiment in Table 5, which shows the depth estimation results when different bias level is used. Unlike previous experiments, Pano3D dataset is not used here for training. As shown in Table 5, the best results are observed when $\rho = 0.1$. These results show that appropriate bias level is important for accurate attention. Meanwhile, based on the results in Table 5, we used $\rho = 0.1$ as default in this paper.

ρ	Abs.rel	Sq.rel	RMS.lin	RMSlog	δ^1	δ^2	δ^3
0.03	0.0347	0.0284	0.2765	0.0941	0.9811	0.9930	0.9963
0.1	0.0338	0.0268	0.2731	0.0933	0.9816	0.9929	0.9963
0.3	0.0352	0.0288	0.2747	0.0942	0.9811	0.9924	0.9964

Table 5: Depth estimation results when different bias level (ρ) is used for 'MMEE-E-EEMM' architecture. Each model is trained and tested with Structured3D dataset [55].

5. Conclusion

In this paper, we propose EGformer for efficient and sophisticated equirectangular depth estimation. The core of EGformer is E(H)V-MSA, which enables to extract local attention in a global manner by considering the equirectangular geometry. To achieve this, we actively utilize the structural prior of EIs when extracting the local attention. Through experiments, we demonstrate that EGformer enables to improve the depth quality level while limiting the computational cost. Considering that EGformer can be generally applied with other attention block as demonstrated in experiments, we expect that EGformer will be extremely beneficial for various 360 vision tasks.

Acknowledgments

This work was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-IT1702-54 and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-00155915, Artificial Intelligence Convergence Innovation Human Resources Development (Inha University)).

References

- [1] Georgios Albanis, Nikolaos Zioulis, Petros Drakoulis, Vasileios Gkitsas, Vladimiro Sterzentsenko, Federico Alvarez, Dimitrios Zarpalas, and Petros Daras. Pano3d: A holistic benchmark and a solid baseline for 360deg depth

- estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3727–3737, 2021. **6, 7, 8**
- [2] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. **6**
- [3] Yoshua Bengio and Jean-Sébastien Senécal. Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Transactions on Neural Networks*, 19(4):713–722, 2008. **5**
- [4] Guy Blanc and Steffen Rendle. Adaptive sampled softmax with kernel based sampling. In *International Conference on Machine Learning*, pages 590–599. PMLR, 2018. **5**
- [5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. **6**
- [6] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. **2**
- [7] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29:730–738, 2016. **6**
- [8] Hsien-Tzu Cheng, Chun-Hung Chao, Jin-Dong Dong, Hao-Kai Wen, Tyng-Luh Liu, and Min Sun. Cube padding for weakly-supervised saliency prediction in 360 videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1420–1429, 2018. **3**
- [9] Sungha Choi, Joanne T Kim, and Jaegul Choo. Cars can’t fly up in the sky: Improving urban-scene segmentation via height-driven attention networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9373–9383, 2020. **2**
- [10] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. In *International Conference on Learning Representations*, 2019. **3**
- [11] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. **2, 3**
- [12] Benjamin Davidson, Mohsan S Alvi, and João F Henriques. 360 camera alignment via segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 579–595. Springer, 2020. **3**
- [13] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022. **1, 2, 3, 4, 7, 8, 9**
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. **1, 2, 3, 8**
- [15] David Eigen, Christian Puhusch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems*, 27, 2014. **6**
- [16] Isaac D Gerg and Vishal Monga. Structural prior driven regularized deep learning for sonar image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021. **2**
- [17] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021. **2**
- [18] Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer. *arXiv preprint arXiv:2209.15001*, 2022. **1**
- [19] Jiayu Jiao, Yu-Ming Tang, Kun-Yu Lin, Yipeng Gao, Jinhua Ma, Yaowei Wang, and Wei-Shi Zheng. Dilateformer: Multi-scale dilated transformer for visual recognition. *IEEE Transactions on Multimedia*, 2023. **1, 3**
- [20] Lei Jin, Yanyu Xu, Jia Zheng, Junfei Zhang, Rui Tang, Shugong Xu, Jingyi Yu, and Shenghua Gao. Geometric structure based and regularized depth estimation from 360 indoor imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 889–898, 2020. **1, 2, 3**
- [21] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*, 2022. **2**
- [22] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. **1**
- [23] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. **1, 2, 4**
- [24] Kevin Matzen, Michael F Cohen, Bryce Evans, Johannes Kopf, and Richard Szeliski. Low-cost 360 stereo photography and video capture. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. **3**
- [25] Greire Payen de La Garanderie, Amir Atapour Abarghouei, and Toby P Breckon. Eliminating the blind spot: Adapting 3d object detection and monocular depth estimation to 360 panoramic imagery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 789–807, 2018. **2, 3**
- [26] Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11536–11545, 2021. **1, 3, 7, 8**

- [27] Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosformer: Rethinking softmax in attention. In *International Conference on Learning Representations*, 2022. 5
- [28] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 2, 6
- [29] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 6
- [30] Ankit Singh Rawat, Jiecao Chen, Felix Xinnan X Yu, Ananda Theertha Suresh, and Sanjiv Kumar. Sampled softmax with random fourier features. *Advances in Neural Information Processing Systems*, 32, 2019. 5
- [31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3
- [32] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoforner: Panorama transformer for indoor 360 depth estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 195–211. Springer, 2022. 1, 2, 3, 7, 8
- [33] Yu-Chuan Su and Kristen Grauman. Learning spherical convolution for fast features from 360 imagery. *Advances in Neural Information Processing Systems*, 30, 2017. 3
- [34] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1047–1056, 2019. 3
- [35] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2573–2582, 2021. 3
- [36] Apoorv Vyas, Angelos Katharopoulos, and François Fleuret. Fast transformers with clustered attention. *Advances in Neural Information Processing Systems*, 33:21665–21674, 2020. 5
- [37] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In *2019 International Conference on 3D Vision (3DV)*, pages 348–357. IEEE, 2019. 6
- [38] Fu-En Wang, Hou-Ning Hu, Hsien-Tzu Cheng, Juan-Ting Lin, Shang-Ta Yang, Meng-Li Shih, Hung-Kuo Chu, and Min Sun. Self-supervised learning of depth and camera motion from 360 videos. In *Asian Conference on Computer Vision*, pages 53–68. Springer, 2018. 3, 6
- [39] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 462–471, 2020. 3, 7, 8
- [40] Fu-En Wang, Yu-Hsuan Yeh, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. Bifuse++: Self-supervised and efficient bi-projection fusion for 360 depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3
- [41] Ning-Hsu Wang, Bolivar Solarte, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. 360sdnet: 360 stereo depth estimation with learnable cost volume. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 582–588. IEEE, 2020. 3
- [42] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 2
- [43] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 2
- [44] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2
- [45] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4794–4803, 2022. 2, 3, 4
- [46] Rui Xu, Xintao Wang, Kai Chen, Bolei Zhou, and Chen Change Loy. Positional encoding as spatial inductive bias in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13569–13578, 2021. 2
- [47] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *Proceedings of the IEEE/CVF International Conference on Computer vision*, pages 16269–16279, 2021. 2
- [48] Jie Yang, Zhiquan Qi, and Yong Shi. Learning to incorporate structure knowledge for image inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12605–12612, 2020. 2
- [49] Shang-Ta Yang, Fu-En Wang, Chi-Han Peng, Peter Wonka, Min Sun, and Hung-Kuo Chu. Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3363–3372, 2019. 3
- [50] Yang Yang and Xiaojie Guo. Generative landmark guided face inpainting. In *Pattern Recognition and Computer Vision: Third Chinese Conference, PRCV 2020, Nanjing, China, October 16–18, 2020, Proceedings, Part I 3*, pages 14–26. Springer, 2020. 2

- [51] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 3
- [52] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021. 2
- [53] Ilwi Yun, Hyuk-Jae Lee, and Chae Eun Rhee. Improving 360 monocular depth estimation via non-local dense prediction transformer and joint supervised and self-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3224–3233, 2022. 1, 3, 6, 7, 8
- [54] Wei Zeng, Sezer Karaoglu, and Theo Gevers. Joint 3d layout and depth prediction from a single indoor panorama image. In *European Conference on Computer Vision*, pages 666–682. Springer, 2020. 3
- [55] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 519–535. Springer, 2020. 1, 6, 7, 8, 9
- [56] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019. 2, 3
- [57] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, Federico Alvarez, and Petros Daras. Spherical view synthesis for self-supervised 360 depth estimation. In *2019 International Conference on 3D Vision (3DV)*, pages 690–699. IEEE, 2019. 3
- [58] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 448–465, 2018. 3, 6