SCCM: Spherically Consistent Coarse Matching for ERP Dense Feature Correspondence

Gyeonggwan Lee[1,2], Eunsoo Im[1], Seunghwan Hong[1], and Junghun Suh[1]

[1] Kakao Mobility Corp., Seongnam, Republic of Korea
[2] Korea University, Seoul, Republic of Korea
{gandan.lee,soo.hd,logan.sh,jude.suh}@kakaomobility.com
Project page: https://gandanlee.github.io/sccm/

Abstract. Equirectangular projection (ERP) is the standard representation for 360° imagery, and robust dense feature matching on ERP underpins panoramic stereo, view synthesis, and omnidirectional SLAM. Dense matchers trained on flat images degrade systematically on ERP because the chart introduces three coupled distortions—topological, metric, and area—that standard coarse matching and visibility estimation do not explicitly model. We show that correcting the three distortions at the coarse-stage interfaces where they arise—pairwise distortions in attention, per-pixel distortion in covisibility gating—improves PCK@1° from 0.229 to 0.275 on Matterport3D under a fixed coarse scaffold, with the refiner architecture unchanged—our central result. Concretely, SCCM (Spherically Consistent Coarse Matching) augments a chart-naïve cross-attention/dual-softmax coarse matcher with two sphere-derived priors: Spherical Positional Attention (SPA) pairs a yaw-periodic RoPE (topology) with a tangent-plane bias (metric), and Area-Aware Covisibility (AAC) applies a pre-sigmoid log-area correction (area). The chart-naïve scaffold serves as a controlled reference, separating the scaffold-replacement effect from the spherical-prior effect. Instantiated in the RoMa V1 framework with the same frozen encoder, refiner architecture, and loss, SCCM also outperforms the ERP-native EDM (0.163) and an ERP-retrained RoMa V1 (0.198) under a unified ERP dense matching protocol, while perspective-trained matchers largely fail on ERP. It further transfers zero-shot to Stanford2D3D and, when trained on outdoor Holo360D, leads there as well.

Keywords: Dense matching · 360° images · Spherical geometry

1 Introduction

Dense feature matching assigns every pixel of one image a corresponding location in another, producing the dense correspondence field that underpins 360° depth, pose, and view synthesis. On 360° content this matching runs on the equirectangular projection (ERP)—the *chart* that unrolls the viewing sphere onto a flat rectangle. Modern dense matchers solve the task in two stages [8, 10, 26]: a *coarse matcher* first proposes low-resolution match anchors, and a *refiner* then

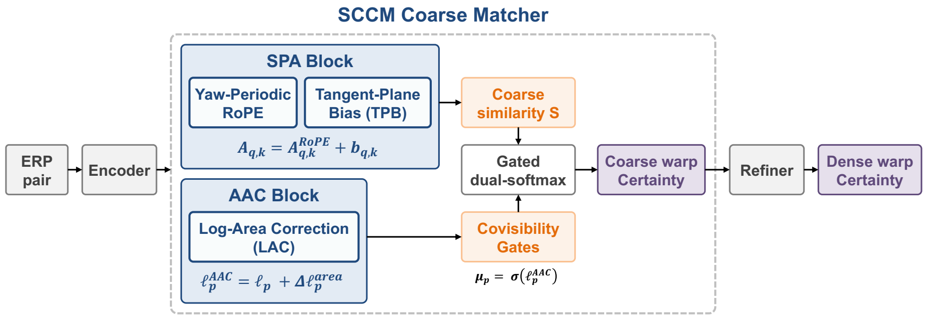

Fig. 1: SCCM coarse matcher: pipeline overview. SCCM injects sphere-derived priors into a fixed chart-naïve coarse scaffold (denoted R1 in Sec. 4), while keeping the encoder, refiner architecture, and loss unchanged. The inherited scaffold consists of cross-/self-attention, covisibility gating, and gated dual-softmax. The *blue* modules are the proposed sphere-aware priors: yaw-periodic RoPE (topology) and Tangent-Plane Bias (metric) in SPA (Fig. 2), and Log-Area Correction (area) in AAC (Fig. 3); *orange* boxes denote coarse matching signals passed to gated dual-softmax, *grey* boxes are inherited unchanged, and *purple* boxes are outputs. The output is a dense warp $W_{A\to B}$ with per-pixel certainty c.

sharpens them into precise correspondences. These matchers achieve strong results on perspective images but degrade systematically on ERP, where the chart's latitude-dependent stretch and longitudinal seam violate the planar assumptions built into their coarse stage.

ERP distortion is not a single artifact but a composite of three computational challenges. **Topology**: the longitudinal $\pm\pi$ seam breaks positional continuity. **Pairwise metric**: a longitudinal stretch by $1/\cos\varphi$ grows toward the poles (here φ is latitude, 0 at the equator and $\pm\frac{\pi}{2}$ at the poles, so $1/\cos\varphi\to\infty$), making Euclidean chart offsets poorly aligned with true geodesic offsets, especially near the poles. **Per-pixel area**: the $\cos\varphi$ scaling of the sphere's area element (the patch of sphere a single pixel covers, which shrinks toward the poles) over-represents polar regions in pixel-uniform sampling and can bias per-pixel matchability estimation toward polar candidates. The first two are *pairwise*—they concern the relation between two feature locations, so we correct them inside *attention* (which scores pairwise interactions between feature locations); the third is *per-pixel*—it concerns whether a single pixel is worth matching at all, so we correct it inside the covisibility *gate* (a per-pixel keep/suppress weight). A single post-hoc adjustment of the final matching cost risks conflating these distinct pairwise and per-pixel effects; we therefore inject each correction at the interface where its distortion arises.

Concretely, we first define a chart-naïve cross-/self-attention + dual-softmax coarse scaffold, then introduce **SCCM** (Spherically Consistent Coarse Matching) as the *same* scaffold augmented with two sphere-aware blocks, SPA and AAC, while keeping the encoder, refiner architecture, and loss unchanged (Fig. 1). SCCM targets the coarse stage, where ERP chart distortions first corrupt the anchors passed to refinement. The **Spherical Positional Attention (SPA)** block augments the attention layers of the coarse cascade with a yaw-periodic Rotary Position Embedding (RoPE; topology) and a tangent-plane bias on the sphere (pairwise metric). The **Area-Aware Covisibility (AAC)** block augments the covisibility gate with a pre-sigmoid log-area correction (per-pixel area). These corrections are modular, sphere-derived priors, yet not independent heuristics: all three follow from a single principle—each ERP distortion is corrected at the coarse-stage decision interface where it enters, pairwise distortions in the attention logits and per-pixel distortion in the covisibility logit—making SCCM a single, spherically consistent set of corrections to coarse matching rather than three unrelated add-ons.

Importantly, part of SCCM's full-system gain comes from the scaffold replacement itself, not from spherical geometry. We therefore use the chart-naïve scaffold as a controlled reference and report the scaffold-replacement and spherical-prior components separately (Sec. 5.2).

In summary, we make three contributions:

- **Distortion-to-interface formulation.** We formulate ERP coarse matching as a distortion-to-interface problem: topology and metric distortions are pairwise and are corrected inside attention, while area distortion is per-pixel and is corrected inside covisibility gating.
- **SPA and AAC.** A spherical attention prior (SPA) combining yaw-periodic RoPE for seam-continuous topology with a tangent-plane bias for geodesic pairwise-metric calibration, and an area-aware covisibility prior (AAC) applying a pre-sigmoid log-area correction to the gate.
- **Controlled evaluation protocol.** An evaluation that separates scaffold gains from spherical-prior gains, with the priors alone improving PCK@1° from 0.229 to 0.275 under a fixed scaffold.

This controlled decomposition is the basis of our evaluation: we report both full-system comparisons and fixed-scaffold ablations to separate architectural replacement from the proposed spherical priors (Sec. 5.3, Tab. 2).

2 Related Work

Perspective and ERP matchers. Perspective matchers—SuperGlue [22], LoFTR [26], DKM [8], and the RoMa family [9, 10]—do not explicitly model ERP chart topology or latitude-dependent metric distortion in their coarse stage, which can leave ERP-specific coarse errors for the refiner to resolve. 3D pointmap regressors (DUSt3R [30], MASt3R [15]) are primarily trained for perspective views. Among ERP methods, sparse spherical matchers (SPHORB [32], SphereGlue [11]) apply geometry only at isolated keypoints, whereas dense matching exposes *every* pairwise matching score to chart distortion. CoMatch [16] introduces covisibility-aware semi-dense matching for perspective images (covisibility-guided token condensing and attention); SCCM instead uses covisibility as a sphere-area-corrected gate for ERP (Sec. 4.2). EDM [14] is the closest prior work: an ERP-native dense matcher with spherical input embeddings and sphere-aware refinement. The difference is one of *where* geometry is injected: EDM injects spherical information at the input and refinement levels, whereas SCCM injects geometry directly into the two generic coarse-stage decision interfaces—the pairwise attention logits and the per-pixel covisibility logits (Sec. 4)—which EDM does not explicitly correct in the same way.

Sphere-aware encodings. Spherical convolutions [4, 5, 27], tangent remapping [7], and panoramic transformers [2, 17] encode geometry within a *single* image and largely leave the relative structure between two feature grids unaddressed, while standard planar RoPE variants used in vision [12, 25] encode 2D chart offsets but do not enforce the 2π longitudinal periodicity required at the ERP seam. Recent panoramic models handle the seam or chart distortion in their positional encodings or tokens: Dense360 [33] folds horizontal indices symmetrically about the image center and rescales them by one global latitude factor; RoPE Rolling [21, 31] shifts the seam to a different longitude per head rather than removing it; SpheRoPE [13] snaps high-frequency harmonics to integers and encodes low frequencies on the sphere, for generation; PanoFormer [23] uses tangent-patch tokens for monocular depth. None is designed for dense matching between two panoramas. SCCM instead builds a seam-periodic pairwise phase and latitude corrections into the coarse attention logits and the covisibility gate.

3 Preliminaries

ERP geometry. An ERP pixel at normalized coordinates $(u, v) \in [-1, 1]^2$ maps to longitude/latitude $(\lambda, \varphi) = (\pi u,\ \frac{\pi}{2} v)$ and thence to a unit viewing ray on the sphere S^2,

$$\mathbf{r}(\varphi, \lambda) = (\cos\varphi \sin\lambda,\ \sin\varphi,\ \cos\varphi\cos\lambda) \in S^2, \tag{1}$$

whose area element is $\cos\varphi$. All SCCM priors derive from this map.

We briefly recall the three standard mechanisms that SCCM modifies: RoPE for relative phase, CPB for additive attention bias, and covisibility gating for filtering non-matchable pixels. This section only fixes notation; the sphere-specific modifications are introduced in Sec. 4.

1. **RoPE.** *Rotary Position Embedding* [25] rotates the query/key vectors by position-dependent phases so that their dot product carries relative-position information. SPA makes the longitudinal phase yaw-periodic across the seam (Sec. 4.1).
2. **CPB.** *Continuous Position Bias* [18] maps a continuous relative coordinate through a small shared MLP to a scalar added to the pre-softmax attention logits. SPA feeds it a spherical log-map offset instead of a planar pixel offset (Sec. 4.1).
3. **Covisibility gating.** Occlusion and limited overlap leave many pixels without a valid match, so dense matchers predict a per-pixel matchability logit ℓ and gate $\mu = \sigma(\ell) \in [0, 1]$ to down-weight non-covisible pixels before *dual-softmax* matching—related notions include CoMatch's covisibility [16],

LoFTR's *confidence* [26], and RoMa's *certainty* [10]. AAC modifies this pre-sigmoid logit with a sphere-area correction (Sec. 4.2).

4 Method

We first define a chart-naïve coarse-matching scaffold, denoted **R1**: cross-/self-attention [22] followed by covisibility-gated dual-softmax [26, 29] (Sec. 3). R1 treats ERP as a flat image—no spherical positional encoding, no tangent-plane metric bias, and no area correction. R1 serves two roles: a strong chart-naïve baseline in its own right (Sec. 5.2), and the fixed scaffold on which spherical priors are isolated (Sec. 5.3). **SCCM** is then R1 augmented with SPA (Sec. 4.1) and AAC (Sec. 4.2). We instantiate this scaffold inside the RoMa V1 [10] framework because its coarse stage is cleanly separable and its training recipe is reproducible: we replace RoMa V1's Gaussian-process coarse stage with R1, keep the DINOv2-Large [19] encoder, the ConvRefiner architecture, and the loss unchanged, freeze the DINOv2 encoder, and train all remaining modules—including the refiner—from random initialization.

By contrast, RoMa V2's coupled architectural changes and the lack of a comparable public recipe at the time of writing would confound the ablation (Supp. Sec. G), so V2 appears only as a zero-shot perspective baseline. The ERP-native EDM [14] is evaluated from its released, *Matterport3D-trained* checkpoint (in-domain; Supp. Sec. G), and we additionally retrain RoMa V1 on ERP under the same protocol as a controlled, same-budget baseline.

Both modules act only at generic coarse-stage interfaces: SPA modifies the attention logits, and AAC the covisibility logits. The additions are lightweight—a $\sim$1.1K-parameter shared MLP for TPB and one scalar per covisibility side for LAC, with a cached inference overhead of about 5% (Supp. Sec. L)—and each is added with a geometry-motivated initialization (TPB zero-initialized as a no-op, LAC at $\alpha_{\mathrm{LAC}}=1$), keeping the cumulative ablation strictly comparable.

4.1 Spherical Positional Attention (SPA)

SPA modifies the relative-position component of the coarse attention layers via two parts (illustrated in Fig. 2): a yaw-periodic RoPE for longitudinal seam consistency, and a tangent-plane bias for the local pairwise metric.

(i) Yaw-Periodic RoPE. On ERP a yaw rotation of the camera (looking left/right about the vertical axis) shifts every longitude $\lambda \to \lambda + \delta$ and wraps across the $\pm\pi$ seam, so a seam-consistent positional encoding must be periodic in λ. Standard planar RoPE [25] uses geometric frequencies ω_i that are not constrained to integer harmonics. Applying them to the ERP longitude $\lambda \in [-\pi, \pi)$ would shift the relative phase $\omega_i(\lambda_q - \lambda_k)$ by $2\pi\omega_i$—generally not a multiple of 2π—when one token's stored longitude wraps across the seam, violating the chart's longitudinal periodicity. Integer harmonics are in fact the only frequencies consistent with this 2π-periodicity (Supp. Sec. H analyzes the resulting frequency resolution). We split each attention head's channels—let d_h be the

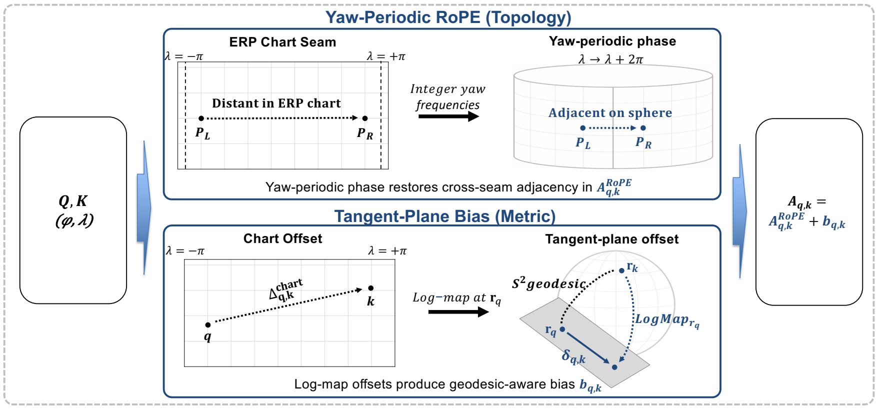

Fig. 2: Spherical Positional Attention (SPA). From coarse query/key tokens Q, K with sphere coordinates (φ, λ), SPA forms two sphere-aware positional terms that merge at the pre-softmax attention logit. **(top)** Yaw-periodic RoPE makes the relative phase strictly 2π-periodic in longitude, so seam-neighbor tokens remain adjacent across the ERP boundary. **(bottom)** Tangent-Plane Bias, instantiating the CPB framework [18] on the sphere, replaces the planar chart offset $\Delta_{q,k}^{\text{chart}}$ with the spherical log-map offset $\boldsymbol{\delta}_{q,k}$ computed in the query tangent plane at $\mathbf{r}_q$, producing a geodesic-aware additive bias $b_{q,k}$. The resulting logit is $A_{q,k} = A_{q,k}^{\text{RoPE}} + b_{q,k}$ (Sec. 4.1).

per-head channel dimension—into two halves of dimension $d_h/2$. The first half encodes latitude with the standard geometric schedule (latitude is bounded, non-periodic), while the second half encodes longitude with integer frequencies $\omega_i^\lambda = i$. For a query at latitude/longitude $(\varphi_q, \lambda_q) \in [-\pi/2, \pi/2] \times [-\pi, \pi)$ and key at (φ_k, λ_k), the longitudinal relative phase becomes

$$\Delta\theta_i^\lambda = i \cdot (\lambda_q - \lambda_k), \qquad i = 1, \ldots, d_h/4, \tag{2}$$

where i ranges up to $d_h/4$ to account for the sin / cos pairings of the $d_h/2$-dimensional longitude half. Because these longitudinal frequencies are integers, a seam wrap $(\lambda \to \lambda + 2\pi)$ shifts each relative phase by a multiple of 2π. This leaves the rotation matrix $\mathbf{R}(\Delta\theta_i^\lambda)$—the standard 2-D RoPE rotation applied to each sin / cos channel pair—unchanged, making the RoPE-induced positional term strictly yaw-periodic. The latitude half retains the standard RoPE form.

(ii) Tangent-Plane Bias (TPB). A relative-position bias is intended to calibrate spatial relations between feature locations; on ERP the chart-pixel offset used by standard biases does not correspond to uniform geodesic offsets across latitude, so a planar bias can favor chart proximity over true spherical proximity. TPB corrects the metric of this bias. It instantiates the Continuous Position Bias (CPB) framework [18] on the sphere: a continuous relative coordinate is mapped to a scalar bias by a shared MLP and added to the pre-softmax attention logits. We adopt CPB's standard integration pattern (pre-softmax additive scalar, shared MLP across queries, output-zero-initialized) and substitute the input coordinate from planar pixel offset to a spherical log-map offset computed

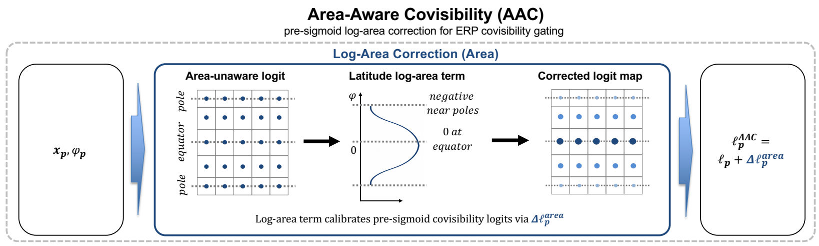

Fig. 3: Area-Aware Covisibility (AAC). Given a per-pixel feature $\mathbf{x}_p$ and latitude φ_p, AAC adds a latitude-dependent log-area term to the covisibility logit before the sigmoid: $\ell_p^{\mathrm{AAC}} = \ell_p + \alpha_{\mathrm{LAC}} \log \max(\cos\varphi_p, \epsilon)$. **(left)** the bare logit $\ell_p = \mathrm{MLP}(\mathbf{x}_p)$ is area-unaware, whereas ERP pixels represent smaller spherical area near the poles; **(middle)** the log-area term is 0 at the equator and negative toward the poles; **(right)** after sigmoid, the corrected logit yields an area-aware gate that down-weights pixel-overrepresented polar candidates before dual-softmax matching, with learnable $\alpha_{\mathrm{LAC}} = \mathrm{softplus}(\alpha_{\mathrm{raw}})$ (one per image side, init $=1$). See Sec. 4.2.

in the query's tangent plane. Let $\mathbf{B}_{\mathbf{r}_q} \in \mathbb{R}^{3\times 2}$ be an orthonormal tangent basis at the query ray $\mathbf{r}_q$ (Eq. (1)), aligned with the ERP latitude/longitude directions; a near-pole-stable construction and antipodal handling are deferred to Supp. Sec. A. For every query we precompute the 2-D geodesic offset of every key via the spherical log-map into this basis:

$$\boldsymbol{\delta}_{q,k} = \mathbf{B}_{\mathbf{r}_q}^{\top} \operatorname{LogMap}_{\mathbf{r}_q}(\mathbf{r}_k) \in \mathbb{R}^2, \tag{3}$$

a spherical log-map offset that replaces the planar pixel offset. It is Fourier-encoded (γ) and mapped by a small shared MLP (zero-initialized output) to a scalar bias added to the pre-softmax logit $A_{q,k}^{\mathrm{RoPE}}$ (the scaled RoPE-rotated query–key dot product):

$$b_{q,k} = \mathrm{MLP}\big(\gamma(\boldsymbol{\delta}_{q,k})\big), \qquad A_{q,k} = A_{q,k}^{\mathrm{RoPE}} + b_{q,k}. \tag{4}$$

Crucially, $\boldsymbol{\delta}_{q,k}$ depends only on the two grid locations' chart coordinates—not on image content or pose—so it is a content-independent spherical prior, precomputed once and shared across heads, SPA layers, and self- and cross-attention; only the MLP and one latitude-scaling scalar are learned (encoding details in Supp. Sec. A).

4.2 Area-Aware Covisibility (AAC)

AAC curbs the tendency of ERP's pixel-uniform sampling to over-represent polar candidates at the covisibility gating stage. Concretely, AAC augments the covisibility head with a single Log-Area Correction (LAC) term defined below; the head structure itself is otherwise identical to the baseline (chart-naïve) covisibility head. The contribution of AAC is not learning a new area function—the ERP area element is known analytically—but *placing* this area prior before the sigmoid, where the model decides whether a pixel should participate in matching. Applying the same factor after dual-softmax, or only in the loss, would not directly calibrate the pre-sigmoid covisibility decision.

Log-Area Correction (LAC). At the covisibility gating stage, the covisibility head produces a per-pixel covisibility logit $\ell = \mathrm{MLP}(\mathbf{x})$, where $\mathbf{x}$ is a learned feature vector. This logit is mapped by a sigmoid to a soft gating probability $\mu = \sigma(\ell)$ used to filter spurious matches before the dual-softmax step. This gate is passed to the refiner as the coarse certainty signal, so area-correcting μ affects the certainty refined downstream.

To correct this polar over-representation, we augment the bare covisibility logit for each pixel feature $\mathbf{x}_p \in \mathbb{R}^d$ at latitude φ_p with a learnable LAC term (Fig. 3). The corrected gate is $\mu_p = \sigma(\ell_p^{\mathrm{AAC}})$, where

$$\ell_p^{\mathrm{AAC}} = \mathrm{MLP}(\mathbf{x}_p) + \alpha_{\mathrm{LAC}} \cdot \log \max\bigl(\cos\varphi_p,\, \epsilon\bigr), \qquad \epsilon = 10^{-3}. \tag{5}$$

The logarithm converts the multiplicative area factor into an additive logit correction, matching the pre-sigmoid interface of the covisibility gate (we abbreviate the two terms as $\ell_p + \Delta\ell_p^{\mathrm{area}}$ in Fig. 3). The clamp ϵ is used only for numerical stability near the poles; sensitivity is reported in Supp. Sec. K. In contrast to input- or refinement-level spherical embeddings, LAC enters the *pre-sigmoid* covisibility logit as a learnable scalar $\alpha_{\mathrm{LAC}} = \mathrm{softplus}(\alpha_{\mathrm{raw}}) > 0$ on the area term (one per image side $v \in \{A, B\}$), initialized at $\alpha_{\mathrm{LAC}} = 1$ so that the correction starts exactly at the analytic ERP-to-sphere area Jacobian $|J| = \cos\varphi$. LAC's contribution is the *placement* of the area term inside the gate, not the tuning of α, which converges near that analytic value (Supp. Sec. K).

5 Experiments

5.1 Experimental Setup

Implementation. All models retrained on Matterport3D share a fixed 1 M-sample protocol (448×896 ERP inputs, frozen DINOv2-Large, original RoMa V1 loss) with the same initialization seed, data split, and evaluator, so observed differences are primarily attributable to the model changes. Each ablation configuration is trained once, with paired bootstrap confidence intervals on the full test split for evaluation-side uncertainty. For training-side variance, the two ablation endpoints were retrained with two more seeds: the chart-naïve→SCCM margin holds on every seed (+4.6/+5.3/+4.9 pp; seed s.d. ≤ 0.6 pp; Supp. Sec. R), whereas the +0.9 pp LAC step is a single-run estimate.

Datasets. We train on Matterport3D [3] (MP3D, indoor) using its official benchmark split, evaluate zero-shot on Stanford2D3D [1] (indoor), and continue training the Matterport3D models on Holo360D [20] (outdoor; Sec. 5.2). All retrained models use $H \times W = 448 \times 896$ ERP inputs (matching the DINOv2 patch grid), with native panoramas resized to this resolution (native dataset resolutions in Supp. Sec. A/B). Released checkpoints are each evaluated at their own native

Table 1: Main full-system comparison on (a) Matterport3D, (b) zero-shot Stanford2D3D, and (c) outdoor Holo360D, where all three matchers are trained on Holo360D from their Matterport3D checkpoints under one protocol. All methods are evaluated end-to-end under the same ERP dense matching metrics. Retrained/controlled rows in (a, b) share the protocol of Sec. 5.1, while external baselines use their released public weights. The mechanism-isolated contribution of the sphere-aware priors, measured as chart-naïve → SCCM under a fixed scaffold, is reported in Tab. 2. Stanford2D3D uses the overlap range $0.30 \leq \text{ov} \leq 0.80$; details in Supp. Sec. B. Best **bold**, second <u>underlined</u>. †SphereGlue [11] is sparse and not directly comparable to dense methods.

Method	Train data	PCK@1°↑	PCK@3°↑	PCK@5°↑	MAE°↓	Med.°↓
(a) Matterport3D test — in-distribution (15,682 pairs)						
Perspective baselines (trained on perspective, zero-shot on ERP)						
RoMa V1 [10]	ScanNet [6] (indoor, persp.)	0.023	0.058	0.081	57.46	49.74
RoMa V2 [9]	Mixed (incl. indoor, persp.)	0.040	0.126	0.201	36.10	18.09
ERP-native baselines (released checkpoints, not retrained)						
SphereGlue† [11]	Custom Synth. (indoor, ERP)	0.028	0.072	0.095	64.50	65.45
EDM [14]	MP3D (indoor, ERP)	0.163	0.400	0.509	16.78	4.79
ERP-retrained baseline						
RoMa V1 (ERP-retrained) [10]	MP3D (indoor, ERP)	0.198	0.538	0.711	6.60	2.76
Fixed-scaffold control (ours, no sphere geometry)						
chart-naïve (cross-attn+DS, no PE)	MP3D (indoor, ERP)	<u>0.229</u>	<u>0.609</u>	<u>0.769</u>	<u>5.68</u>	<u>2.23</u>
Ours						
SCCM (full)	MP3D (indoor, ERP)	**0.275**	**0.665**	**0.806**	**5.36**	**1.88**
(b) Stanford2D3D test — out-of-distribution / zero-shot (8,744 pairs)						
Perspective baselines (trained on perspective, zero-shot on ERP)						
RoMa V1 [10]	ScanNet (indoor, persp.)	0.022	0.051	0.072	64.08	58.99
RoMa V2 [9]	Mixed (incl. indoor, persp.)	0.043	0.125	0.201	42.15	18.59
ERP-native baselines (released checkpoints, not retrained)						
SphereGlue† [11]	Custom Synth. (indoor, ERP)	0.034	0.072	0.093	67.96	62.78
EDM [14]	MP3D (indoor, ERP)	0.104	0.262	0.355	35.70	10.69
ERP-retrained baseline						
RoMa V1 (ERP-retrained) [10]	MP3D (indoor, ERP)	0.167	0.452	<u>0.628</u>	<u>17.80</u>	3.45
Fixed-scaffold control (ours, no sphere geometry)						
chart-naïve (cross-attn+DS, no PE)	MP3D (indoor, ERP)	<u>0.179</u>	<u>0.475</u>	0.627	18.86	<u>3.26</u>
Ours						
SCCM (full)	MP3D (indoor, ERP)	**0.229**	**0.564**	**0.697**	**17.09**	**2.45**
(c) Holo360D test — outdoor, in-the-wild (8,000 pairs)						
Trained under one protocol (ours; Supp. Sec. P)						
RoMa V1 (ERP-retrained) [10]	Holo360D (outdoor)	0.322	0.592	0.726	5.66	2.12
chart-naïve (cross-attn+DS, no PE)	Holo360D (outdoor)	<u>0.331</u>	<u>0.626</u>	<u>0.761</u>	<u>5.02</u>	<u>1.93</u>
SCCM (full)	Holo360D (outdoor)	**0.357**	**0.645**	**0.776**	**4.94**	**1.76**

resolution, with ground truth and metrics on that grid (resolution-matched control for EDM in Supp. Sec. R). SCCM assumes gravity-aligned (upright) ERP, as in the Matterport3D and Stanford2D3D benchmarks; robustness to camera tilt is analyzed in Supp. Sec. F.

Evaluation metrics. We evaluate angular error on the sphere, $\theta = \arccos(\mathbf{r}_{\text{pred}} \cdot \mathbf{r}_{\text{gt}})$, rather than pixel-space endpoint error (biased by ERP's $\cos\varphi$ anisotropy), and report PCK (the fraction of valid pixels below angular thresholds) at $\{1, 3, 5\}^\circ$ with mean (MAE) and median angular error; PCK@1° is the primary strict-precision metric.

Full training protocol, dataset construction, and complexity details are in Supp. Secs. A, B, and L.

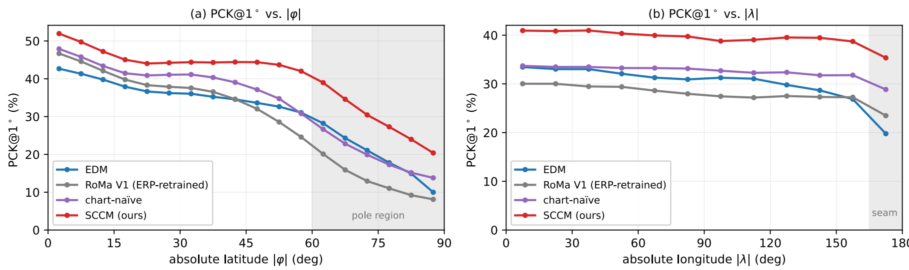

Fig. 4: Precision analysis on Matterport3D test (full test set, 15,682 pairs). (a) PCK@1° by absolute latitude $|\varphi|$: SCCM is consistently highest across latitude bands, with a larger margin toward the poles (shaded). (b) PCK@1° by absolute longitude $|\lambda|$: SCCM shows a smaller drop near the ERP seam ($|\lambda| = 180°$, shaded) than EDM, while the chart-naïve scaffold lies between EDM and SCCM.

5.2 Benchmark Comparison

Model groups. Under the shared evaluation protocol of Sec. 5.1 we compare: released *perspective baselines* (RoMa V1 [10], RoMa V2 [9], zero-shot on ERP), *ERP-native baselines* (sparse SphereGlue [11], dense EDM [14]), an *ERP-retrained RoMa V1*, our *chart-naïve* scaffold (cross-attention + dual-softmax, no positional encoding), and *SCCM* (chart-naïve + SPA + AAC). RoMa V1 thus plays three roles: zero-shot checkpoint, ERP-retrained baseline, and the base architecture for our controlled study. The ablation (Sec. 5.3) adds two mechanism-only positional-encoding (PE) controls on that scaffold—standard geometric-frequency RoPE and EDM's input-side absolute PE—re-implemented as the mechanism only, not the full EDM system. EDM is evaluated from its released checkpoint—its training code is not public (Supp. Sec. G)—so that comparison is matched at evaluation but not at training recipe. Four further matchers (DKM, LoFTR, MASt3R, VGGT [28]) evaluated zero-shot on ERP all fall below EDM (Supp. Sec. Q).

Decomposing the gain. We report two distinct gains on Matterport3D (Tab. 1a): a *scaffold gain* from replacing RoMa V1's GP coarse stage with our chart-naïve R1 (0.198 → 0.229 PCK@1°, +3.1 pp), and a *spherical-prior gain* from adding the three sphere-derived modules on the fixed R1 scaffold (0.229 → 0.275, +4.6 pp; Sec. 5.3). The latter is the mechanism-isolated contribution of SCCM and the central claim of this paper; the full-system numbers below combine both. A paired bootstrap (10,000 resamples) supports this gain: the chart-naïve→SCCM ΔPCK@1° CI excludes zero on MP3D (+4.6 pp) and zero-shot Stanford2D3D (+5.0 pp; Supp. Sec. D).

In-distribution: Matterport3D test. Tab. 1(a) reports results on the official Matterport3D test split. Zero-shot perspective baselines (RoMa V1, RoMa V2) stay below 0.05 PCK@1°, reflecting ERP distortion compounded by the perspective-to-ERP domain shift. Among prior published ERP methods, the ERP-native EDM is strongest (PCK@1° = 0.163, MAE = 16.78°; the sparse

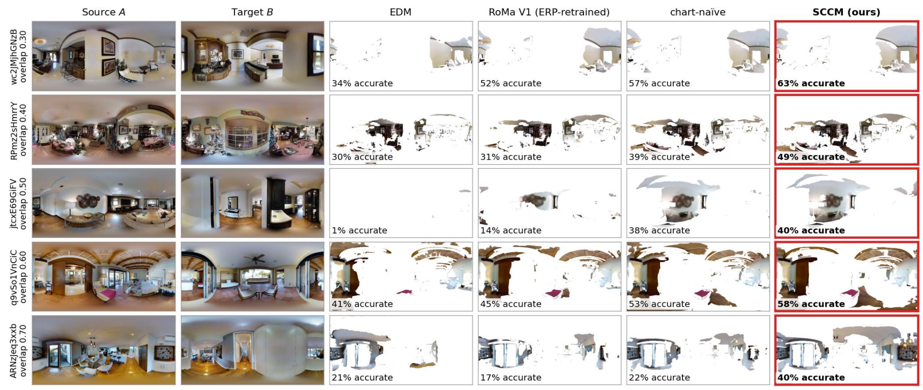

Fig. 5: Qualitative comparison on Matterport3D test. Each method panel shows the predicted warp masked by correctness: white denotes inaccurate or non-covisible pixels. The reported percentage is the fraction of GT-covisible pixels with angular error below 1°. SCCM recovers larger accurate regions than the dense baselines (EDM, RoMa V1 (ERP-retrained)) and the chart-naïve scaffold across representative overlap levels, including high-latitude ceiling/floor regions.

SphereGlue is not directly comparable), and our ERP-retrained RoMa V1 (0.198) already surpasses it. Our chart-naïve backbone surpasses EDM *without any sphere-aware encoding* (0.229 vs. 0.163), and positional-encoding controls on it change little (Tab. 2): it is a strong matching substrate, and generic positional encodings alone do not explain the SCCM gain. SCCM reaches PCK@1° = 0.275, +11.2 pp over EDM and +4.6 pp over the chart-naïve backbone (Tab. 1a). Fig. 5 shows the gap qualitatively across overlap levels.

Absolute vs. pairwise positional encoding. The EDM absolute PE control re-implements EDM's input-side absolute sphere-coordinate embedding on the chart-naïve backbone, testing whether providing sphere coordinates to a strong planar backbone suffices. It performs slightly below the backbone alone (Tab. 2), so SCCM's lead is essentially unchanged against it. Absolute encoding assigns coordinates to each location independently and carries no relative query–key geometry; the gain instead comes from pairwise spherical structure, with the largest step at R1→R2a (+3.8 pp). This control tests the input-side absolute-PE mechanism under the same scaffold, not EDM as a full system.

Out-of-distribution: Stanford2D3D. Tab. 1(b) reports zero-shot evaluation on the filtered Stanford2D3D test-pair set (8,744 pairs). Our numbers differ from the original EDM paper because we use overlap $0.30 \leq \text{ov} \leq 0.80$ instead of > 0.50, which adds harder low-covisibility pairs and excludes near-duplicate ones (Supp. Sec. B). SCCM exceeds every baseline (Tab. 1b), so the sphere-aware advantage transfers across panoramic datasets.

Outdoor. For Tab. 1(c), RoMa V1 (ERP-retrained), chart-naïve, and SCCM are trained from their MP3D checkpoints under one protocol on Holo360D [20], an in-the-wild handheld LiDAR+360° dataset (2.2× deeper than MP3D, median relative tilt 12°, i.e. tilted training data, cf. Supp. Sec. F; scene-disjoint 5/3/4 splits, 8,000 test pairs). SCCM leads on every metric, by +2.6/+3.5 pp PCK@1° over chart-naïve/RoMa V1; zero-shot outdoor transfer is analyzed in Supp. Sec. P.

Latitude-wise behavior. Fig. 4(a) reports PCK@1° by absolute latitude on Matterport3D test. Baselines degrade steeply as $|\varphi|$ grows, since the same chart offset spans different geodesic distances across latitudes. SCCM improves PCK@1° at every latitude, with the margin growing toward the poles ($\approx 1.5\times$ the chart-naïve scaffold at the highest band vs. $\approx 1.1\times$ at the equator), and remains more robust than EDM across the longitudinal seam (Fig. 4b). Per-band decomposition attributes these effects mainly to yaw-periodic RoPE, with TPB adding MAE correction at low and mid latitudes (Supp. Sec. M). Because pixel-uniform evaluation over-represents polar pixels, we also weight pixels by $\cos\varphi$; the chart-naïve → SCCM improvement persists on every metric (PCK@1° +3.9 pp area-weighted vs. +4.6 pp; Supp. Sec. E).

5.3 Ablation Study

Tab. 2 isolates the sphere-aware modules on the fixed chart-naïve scaffold R1 (cross-attention + dual-softmax), whose standing against all baselines is established in Sec. 5.2. R2a–R3 progressively add yaw-periodic RoPE, TPB, and AAC with all other settings held constant, so the +4.6 pp PCK@1° gain from R1 to R3 isolates the spherical modules from the scaffold replacement (full MP3D test set, 15,682 pairs, ∼3.0B pixels per checkpoint).

Module contributions. The three modules act along distinct axes (Tab. 2 lists marginal changes and primary roles; per-band decomposition in Supp. Sec. M). *Yaw-periodic RoPE* supplies most of the precision gain—+3.8 pp PCK@1° over

Table 2: Controlled ablation on the fixed chart-naïve scaffold (Matterport3D test). R2a–R3 cumulatively add yaw-periodic RoPE, TPB, and LAC/AAC to the no-prior scaffold R1; the two § rows are positional-encoding controls on R1. Values in parentheses on PCK@1° and MAE denote the marginal change from the previous chain row (PCK in pp, MAE in degrees), revealing the modules' functional roles: yaw-periodic RoPE drives strict precision, TPB reduces the angular-error tail, and LAC adds area-gated precision. Chain marginals on PCK@1° fold the TPB×LAC interaction into the final row (factorial decomposition in Supp. Sec. M). Protocol as in Sec. 5.1. Best in **bold**.

	PCK° (↑)			**Error° (↓)**		
Row	**@1**	**@3**	**@5**	**MAE**	**Med.**	**Primary role**
(R1) chart-naïve (no PE)	0.229	0.609	0.769	5.68	2.23	planar scaffold
ctrl: EDM abs. PE§	0.224	0.593	0.756	5.97	2.31	*(input-PE control)*
ctrl: standard RoPE§	0.236	0.613	0.764	6.79	2.17	*(rel-PE control)*
(R2a) + yaw-periodic RoPE	0.267 (+3.8)	0.653	0.796	5.87 (+0.19)	1.93	strict precision
(R2b) + TPB (full SPA)	0.266 (−0.1)	0.652	0.799	5.58 (−0.29)	1.95	error-tail (MAE)
(R3) + LAC ≡ SCCM	**0.275** (+0.9)	**0.665**	**0.806**	**5.36** (−0.22)	**1.88**	area-gated precision

R1, consistently positive across longitude bands, including near the seam—and is specific to the 2π-periodic design: substituting standard geometric-frequency RoPE (Tab. 2, §) yields only +0.7 pp and *degrades* MAE ($5.68^\circ \to 6.79^\circ$), supporting longitudinal periodicity as the key factor rather than generic relative encoding. *TPB* leaves strict PCK essentially unchanged (-0.1 pp) while reducing the angular-error tail (MAE $5.87^\circ \to 5.58^\circ$; tail quantiles in Supp. Sec. M); its contribution is metric calibration—conditioning the pre-softmax bias on the true geodesic offset—complementary to the precision gain of yaw-periodic RoPE. *LAC (AAC)* adds the log-area term $\alpha_{\mathrm{LAC}} \log \max(\cos\varphi, \epsilon)$ to the pre-sigmoid covisibility logit, calibrating pixel-overrepresented polar candidates through the known area prior: it lowers MAE (-0.22°) and median (-0.07°) and lifts PCK@1° by +0.9 pp (Tab. 2). A 2×2 factorial with a RoPE+LAC control (Supp. Sec. M) attributes this strict-precision step to the TPB×LAC *interaction* rather than to LAC alone: neither prior improves PCK@1° in isolation, whereas MAE improves additively—consistent with strict-precision gains emerging only when the area-corrected gate is paired with the geodesic pairwise bias. The learned α_{LAC} settles near the analytic $\cos\varphi$ Jacobian; LAC's contribution is the *placement* of the area-correction term, not the tuning of its scalar (an inference-time $\alpha = 1$ override leaves all metrics unchanged; Supp. Sec. K).

Coarse-stage leverage. Coarse-anchor quality is thus the key leverage point under this scaffold: the full +4.6 pp PCK@1° gain is obtained by changing only the coarse-stage modules, with the encoder, refiner architecture, loss, and training protocol fixed. A refinement-stage diagnostic on the full test split (Supp. Sec. O) supports this: SCCM's advantage is already present at the coarse-anchor stage (+4.1 pp) and remains stable through refinement (+4.6 pp at the final output), while the unchanged refiner gives a similar coarse-to-final lift for R1 and SCCM (+10.7 vs. +11.2 pp), so the ERP-specific gain enters through coarse-anchor formation rather than being created by the refiner.

5.4 Downstream Tasks

We evaluate whether the dense matching gains translate to downstream geometry on Matterport3D under a single unified evaluator (Tab. 3). Both tasks use each matcher's per-pixel certainty: pose samples ray matches by certainty, and reconstruction applies one fixed threshold ($\tau = 0.5$) to every method. For **relative pose** (essential-matrix + RANSAC on certainty-weighted ray matches), SCCM improves Pose AUC at every threshold over the chart-naïve scaffold (Tab. 3). Replacing the estimator with the 360-8PA solver of Solarte *et al*. [24] moves AUC@5° by ≤ 0.01 pp for every method (Supp. Sec. R). For **3D reconstruction**, we triangulate under the GT pose to isolate correspondence quality from pose-estimation errors; SCCM leads on *all* seven accuracy, completeness, and F-score metrics, ahead of the chart-naïve scaffold, RoMa V1 (ERP-retrained), and EDM. Ungated reconstruction (Supp. Sec. N) shows the same trend except for mean accuracy; details and top-down visualizations in Supp. Sec. C.

Table 3: Downstream geometry on Matterport3D under a unified evaluator. We report relative-pose AUC and certainty-gated reconstruction metrics. Reconstruction triangulates matches under the GT pose with a fixed certainty gate ($\tau = 0.5$) for all methods; both tasks are evaluated on the 11,574 test pairs where every method yields ≥ 200 confident matches; details in Supp. Sec. C. Best **bold**, second <u>underlined</u>.

	Pose AUC ↑			**Acc** (m) ↓		**Comp** (m) ↓		**Recon. F** ↑		
Method	5°	10°	20°	Mean	Med.	Mean	Med.	@5	@10	@20
EDM [14]	7.97	17.86	29.51	1.452	0.470	0.685	0.336	14.5	27.5	42.7
RoMa V1 (retr.) [10]	11.12	27.26	47.42	0.884	0.294	0.573	0.250	16.0	30.7	48.2
chart-naïve	<u>13.66</u>	<u>31.50</u>	<u>51.76</u>	<u>0.787</u>	<u>0.267</u>	<u>0.512</u>	<u>0.236</u>	<u>17.4</u>	<u>32.7</u>	<u>50.6</u>
SCCM (ours)	**17.41**	**35.84**	**55.09**	**0.772**	**0.247**	**0.496**	**0.218**	**19.3**	**35.5**	**53.2**

6 Conclusion

We introduced SCCM, a coarse matcher that corrects the three geometric distortions of equirectangular projection—topology, pairwise metric, and per-pixel area—in the coarse-stage operations where they arise (attention and covisibility gating), through lightweight modules that leave the encoder, refiner, and loss unchanged. Under our unified ERP dense matching protocol, SCCM outperforms all tested perspective-trained, ERP-retrained, and ERP-native baselines on Matterport3D, transfers zero-shot to Stanford2D3D, and leads on outdoor Holo360D when trained on it. Both modules attach to generic attention and covisibility logits, so they apply to attention-based matchers such as LoFTR; GP-based coarse stages (DKM, original RoMa V1) admit only AAC (Supp. Sec. G).

Limitations and future work. SCCM assumes gravity-aligned ERP and does not explicitly handle camera tilt: under synthetic camera-pitch perturbations all tested ERP matchers degrade sharply (SCCM PCK@1°: $0.273 \rightarrow 0.142$ at 10° pitch, 0.034 at 30°, though still best at every angle; Supp. Tab. 5), and SCCM is not tilt-equivariant (failure case: Supp. Sec. N). It also inherits the frozen encoder's (DINOv2-Large) limitations in texture-less, photometrically ambiguous, or low-overlap regions. SCCM has been trained only on Matterport3D (indoor) and Holo360D (outdoor); applied zero-shot to a further outdoor corpus, the Holo360D-trained model leads only at $\leq 1°$ (Supp. Sec. P). SCCM leaves the refiner sphere-naïve (Supp. Sec. O). Rotation-equivariant coarse matching on $SO(3)$, ambiguity-aware encoders, and sphere-aware fine refinement are future work.

Acknowledgement. This work was supported by Kakao Mobility Corp. and Korea University. We thank the anonymous reviewers and the area chair for their constructive feedback, and the authors of RoMa, EDM, Matterport3D, Stanford2D3D, and Holo360D for publicly releasing their code, models, and data.

References

1. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2D-3D-Semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105 (2017)
2. Benny, Y., Wolf, L.: SphereUFormer: A U-shaped transformer for spherical 360 perception. In: CVPR (2025)
3. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D data in indoor environments. In: 3DV (2017)
4. Cohen, T.S., Geiger, M., Köhler, J., Welling, M.: Spherical CNNs. In: ICLR (2018)
5. Coors, B., Condurache, A.P., Geiger, A.: SphereNet: Learning spherical representations for detection and classification in omnidirectional images. In: ECCV (2018)
6. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: CVPR (2017)
7. Eder, M., Shvets, M., Lim, J., Frahm, J.M.: Tangent images for mitigating spherical distortion. In: CVPR (2020)
8. Edstedt, J., Athanasiadis, I., Wadenbäck, M., Felsberg, M.: DKM: Dense kernelized feature matching for geometry estimation. In: CVPR (2023)
9. Edstedt, J., Nordström, D., Zhang, Y., Bökman, G., Astermark, J., Larsson, V., Heyden, A., Kahl, F., Wadenbäck, M., Felsberg, M.: RoMa v2: Harder better faster denser feature matching. In: ECCV (2026)
10. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: RoMa: Robust dense feature matching. In: CVPR (2024)
11. Gava, C., Mukunda, V., Habtegebrial, T., Raue, F., Palacio, S., Dengel, A.: SphereGlue: Learning keypoint matching on high resolution spherical images. In: CVPRW (2023)
12. Heo, B., Park, S., Han, D., Yun, S.: Rotary position embedding for vision transformer. In: ECCV (2024)
13. Hirschorn, O., Olender, A., Alshan, E., Ideses, I., Fritz, L., Benaim, S.: SpheRoPE: Zero-shot optimization-free 360 panorama generation with spherical RoPE. arXiv preprint arXiv:2606.32033 (2026)
14. Jung, D., Choi, J., Lee, Y., Jeong, S., Lee, T., Manocha, D., Yeon, S.: EDM: Equirectangular projection-oriented dense kernelized feature matching. In: CVPR (2025)
15. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3D with MASt3R. In: ECCV (2024)
16. Li, Z., Lu, Y., Tang, L., Zhang, S., Ma, J.: CoMatch: Dynamic covisibility-aware transformer for bilateral subpixel-level semi-dense image matching. In: ICCV (2025)
17. Ling, Z., Xing, Z., Zhou, X., Cao, M., Zhou, G.: PanoSwin: A pano-style swin transformer for panorama understanding. In: CVPR (2023)
18. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., Guo, B.: Swin transformer V2: Scaling up capacity and resolution. In: CVPR (2022)
19. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
20. Ou, J., Cao, Z., Ren, Y., Li, Z., Zhu, J., Hua, T., Zhang, S., Xiong, H., Zhao, W.: Holo360D: A large-scale real-world dataset with continuous trajectories for advancing panoramic 3D reconstruction and beyond. arXiv preprint arXiv:2604.22482 (2026)

21. Ren, J., Xiang, M., Zhu, J., Dai, Y.: PanoSplatt3R: Leveraging perspective pretraining for generalized unposed wide-baseline panorama reconstruction. In: ICCV (2025)
22. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: CVPR (2020)
23. Shen, Z., Lin, C., Liao, K., Nie, L., Zheng, Z., Zhao, Y.: PanoFormer: Panorama transformer for indoor 360° depth estimation. In: ECCV (2022)
24. Solarte, B., Wu, C.H., Lu, K.W., Sun, M., Chiu, W.C., Tsai, Y.H.: Robust 360-8PA: Redesigning the normalized 8-point algorithm for 360-FoV images. In: ICRA (2021)
25. Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., Liu, Y.: RoFormer: Enhanced transformer with rotary position embedding. arXiv preprint arXiv:2104.09864 (2021)
26. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. In: CVPR (2021)
27. Tateno, K., Navab, N., Tombari, F.: Distortion-aware convolutional filters for dense prediction in panoramic images. In: ECCV (2018)
28. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: VGGT: Visual geometry grounded transformer. In: CVPR (2025)
29. Wang, Q., Zhang, J., Yang, K., Peng, K., Stiefelhagen, R.: MatchFormer: Interleaving attention in transformers for feature matching. In: ACCV (2022)
30. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: DUSt3R: Geometric 3D vision made easy. In: CVPR (2024)
31. Xu, J., Lin, Z., Sun, J., Yang, Y., Luo, Y.: EAGLE-360: Embodied active global-to-local exploration in 360°. arXiv preprint arXiv:2607.02479 (2026)
32. Zhao, Q., Feng, W., Wan, L., Zhang, J.: SPHORB: A fast and robust binary feature on the sphere. IJCV **113**, 143–159 (2015)
33. Zhou, Y., Zhang, T., Zhang, D., Ji, S., Li, X., Qi, L.: Dense360: Dense understanding from omnidirectional panoramas. arXiv preprint arXiv:2506.14471 (2025)