SchellingFormer: Laplacian Matrix-guided Geometric Transformer for Robust Schelling Point Detection

Yihao Chen[1*], Haobo Jiang[1*], Liang Yu[2], Jianmin Zheng[1†]

[1]College of Computing and Data Science, Nanyang Technological University, Singapore
[2]Alibaba Group
{yihao003, haobo.jiang, asjmzheng}@ntu.edu.sg, liangyu.yl@alibaba-inc.com

Abstract

Detecting Schelling Points—salient 3D mesh landmarks that serve as natural reference points for shape analysis—is a challenging problem in geometry processing. While existing CNN-based methods struggle with limited receptive fields and poor geometric context modeling, this paper proposes *SchellingFormer*, a novel Laplacian matrix-guided Geometric Transformer that effectively captures long-range dependencies and discriminative geometric features for robust Schelling point prediction. Our framework consists of two key components: (i) a hybrid geometric feature embedding module that integrates handcrafted descriptors (coordinates, Gaussian curvature, and curvature differences) to encode local geometry, and (ii) a Laplacian-driven vector attention mechanism, where spatial relationships encoded by the Laplacian matrix guide feature aggregation with the Transformer. This approach enables adaptive, geometry-aware message passing and contextual representation learning. Extensive experiments demonstrate that SchellingFormer outperforms state-of-the-art methods across multiple evaluation metrics. Our work bridges the gap between spectral mesh analysis and Transformer-based learning, offering a powerful tool for 3D shape understanding tasks such as shape matching and saliency detection.

Introduction

Schelling points on 3D meshes refer to salient or intuitively "natural" points on a shape that multiple humans are likely to independently agree upon when asked to identify important or special locations (Chen et al. 2012). This concept originates from game theory, where a Schelling point is a solution that people tend to choose by default in the absence of communication, due to its salience or naturalness. Detecting Schelling points thus plays an important role in geometric processing and shape analysis with various applications, including shape recognition (Su et al. 2015), object segmentation (Kaick et al. 2014; Shu et al. 2022a), mesh simplification (Potamias, Ploumpis, and Zafeiriou 2022) and 3D matching (Jiang et al. 2021, 2023a,b). However, owing to the inherent reliance of Schelling points on human perception, combined with real-world challenges such as surface noise, varying mesh resolution, irregular topology, and the absence of explicit semantic cues, accurately and robustly detecting Schelling points remains a significant challenge.

Considerable research has been dedicated to the detection of Schelling points on 3D meshes. Early approaches commonly rely on user studies to capture human consensus in Schelling point selection, thereby providing perceptual ground truth for evaluation (Kim et al. 2010; Chen et al. 2012; Dutağacı, Cheung, and Godil 2012). However, these methods are limited by the need for tedious manual annotation and often depend heavily on local geometric cues such as Gaussian curvature. Recent Schelling point detection typically adopts a two-stage pipeline: first computing saliency scores for mesh elements (e.g., vertices or faces), and subsequently identifying Schelling points based on the saliency distribution. Many methods have been developed to estimate mesh saliency by leveraging local, handcrafted geometric features such as local curvature or multi-scale contrast (Limper, Kuijper, and Fellner 2016; Nouri, Charrier, and Lézoray 2015; Lee, Varshney, and Jacobs 2005; Song et al. 2014; Gal and Cohen-Or 2006; Kim and Varshney 2006; Shilane and Funkhouser 2007; Leifman, Shtrom, and Tal 2016; Wu et al. 2013; Song et al. 2018). While effective to some extent, these methods often fail to capture global structural context that is crucial for perceptual saliency. With advances in deep learning, data-driven approaches have emerged as a popular alternative for Schelling point detection. Several studies demonstrate that advanced neural models can produce promising saliency predictions by learning discriminative geometric features from annotated datasets (Wang et al. 2015; Chen et al. 2022; Shu et al. 2022b, 2024). Nevertheless, these deep-learning-based methods commonly employ CNN-based backbones for feature extraction, which still suffer from limited receptive fields and insufficient geometric context modeling, resulting in reduced prediction accuracy, particularly for objects with complex shapes.

This paper proposes *SchellingFormer*, a novel Laplacian matrix-guided Transformer architecture that models long-range geometric dependencies to learn context-aware geometric representations for robust Schelling point detection. Beyond the limited receptive fields of existing CNN-based approaches, SchellingFormer leverages the global attention mechanism of Transformers, guided by spectral geometric

[*]These authors contributed equally.
[†]Corresponding Author.

priors, to capture both fine-grained local structure information and global geometric contextual structures. Specifically, our SchellingFormer primarily consists of two components: a hybrid geometric feature embedding module and a Laplacian matrix-driven vector attention module. In the feature embedding module, we integrate a set of handcrafted geometric descriptors, including vertex coordinates, Gaussian curvature and curvature difference, to encode local shape characteristics of the mesh. These features serve as the initial input tokens to the Transformer and provide a rich representation of local surface geometry. In the Laplacian matrix-driven vector attention module, we innovatively incorporate the spectral geometric priors into the attention process. Here, we compute the mesh Laplacian to capture intrinsic spatial relationships between neighboring vertices, and use its entries to modulate the self-attention maps. In particular, we employ a vector self-attention mechanism that adaptively learns to map each entry of the Laplacian matrix to attention vectors, enabling adaptive message passing that is explicitly aware of the underlying point cloud structure. Finally, the resulting contextual geometric representations are subsequently passed through a multi-layer perceptron (MLP) to generate a saliency map, from which Schelling points are inferred using our Schelling point extraction module. In summary, our main contributions are as follows:

- We present SchellingFormer, a novel Laplacian matrix-guided Geometric Transformer architecture for robust Schelling point detection. Unlike conventional CNN-based approaches with limited receptive fields, SchellingFormer effectively captures long-range geometric dependencies and learns discriminative, context-aware representations, significantly improving the prediction accuracy.
- We construct a hybrid feature embedding by integrating handcrafted geometric descriptors, including vertex coordinates, Gaussian curvature and curvature differences, to capture detailed local shape information. These features serve as the Transformer's input tokens, enriching the representation of local surface geometry.
- We introduce a novel Laplacian matrix-guided vector attention module that incorporates the Laplacian cues as a spectral geometric prior for geometry-enhanced long-range relationship modeling. By mapping Laplacian entries to attention vectors, this module enables flexible, geometry-aware message passing that effectively captures contextual geometric cues across the object surface.

Extensive experiments on benchmark datasets demonstrate that our method outperforms state-of-the-art techniques in Schelling point detection across multiple evaluation metrics, highlighting its effectiveness and robustness.

Related Work

3D Salient Region Prediction. Identifying semantically meaningful regions is important in 3D applications including computer graphics, virtual reality and cognitive psychology. Early methods (Yee, Pattanaik, and Greenberg 2001; Mantiuk, Myszkowski, and Pattanaik 2004) analyzed 3D saliency by projecting meshes onto 2D domains. A key advancement came from Lee et al. (Lee, Varshney, and Jacobs 2005), who introduced mesh saliency as a surface-based metric using Gaussian-weighted center-surround filters applied to curvature fields. The relevance of 3D saliency is evident in downstream tasks such as mesh segmentation (Katz and Tal 2003; Katz, Leifman, and Tal 2005; Ji et al. 2006; Liu and Zhang 2007; Golovinskiy and Funkhouser 2008; Kaick et al. 2014; Shu et al. 2019, 2022a), descriptor extraction (Gal and Cohen-Or 2006; Castellani et al. 2008; Lian et al. 2011), shape simplification (Menzel and Guthe 2010), and enhancement (Kim and Varshney 2006).

Point Cloud-Based Key Point Detection. A subset of deep learning methods for 3D saliency detection centers on point cloud representations, wherein salient keypoints are extracted from sampled 3D points. USIP (Li and Lee 2019) improves keypoint localization and repeatability by leveraging Feature Pyramid Networks (FPN) and introducing a probabilistic Chamfer loss. UKPGAN (You et al. 2022) proposes a keypoint significance distribution via a dedicated detection network and employs an adversarial loss to jointly enforce sparsity and geometric reconstruction. D3Feat (Bai et al. 2020) integrates a fully convolutional feature descriptor with a salient point detector to achieve robust and accurate keypoint identification. In a complementary direction, Fernández et al. (Fernandez-Labrador et al. 2020) explore unsupervised keypoint learning from collections of misaligned objects belonging to unknown categories. Jakab et al. (Jakab et al. 2021) proposed an unsupervised stable interest point detector capable of identifying highly repeatable and accurately localized keypoints.

Data-Driven Detection of Schelling Points. Due to limited availability of annotated training data, earlier data-driven methods for 3D point-of-interest detection primarily relied on hand-crafted geometric descriptors to estimate the likelihood of each point being a Schelling point. Classical approaches to 3D keypoint detection (Novatnack 2007; Castellani et al. 2008; Sipiran and Bustos 2011) heavily depended on such manually designed geometric features. With the recent progress in deep learning, data-driven techniques have become increasingly prominent in this domain (Sung et al. 2018; Zhu et al. 2023; He et al. 2020). Wei et al. (Wei et al. 2021) proposed a multi-task learning framework that integrates point-to-keypoint offsets with a confidence map, enabling robust estimation of 3D keypoint saliency and correspondences. Shu et al. (Shu et al. 2022b) introduced a projective neural network-based approach that first projects labeled 3D shapes into multiple 2D views, and subsequently learns salient features in an end-to-end manner from these projections. Building upon this work, Shu et al. (Shu et al. 2024) further proposed a coarse-to-fine multi-modal framework for salient point detection, improving precision by leveraging complementary geometric and appearance cues.

SchellingFormer

Problem Definition. Schelling points represents the reference points on 3D surface meshes that are consistently perceived by human observers as perceptually significant, as illustrated in Figure 1. These points often correspond to

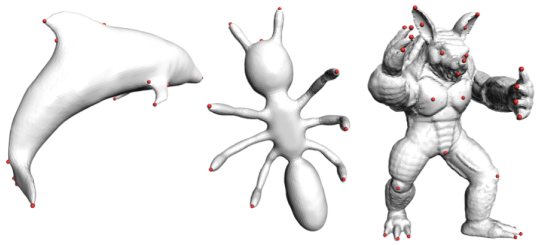

Figure 1: Schelling points collected from human participants, following the protocol in (Chen et al. 2012).

regions that exhibit richer geometric or semantic content compared to their surroundings. Notably, their number and spatial distribution can vary considerably across different meshes, even among instances within the same object category. The problem considered in this paper is to infer these perceptually salient points $\{\mathbf{p}_i^s \in \mathbb{R}^3\}$ from a given 3D surface mesh $\mathcal{M} = \{\mathcal{F}, \mathcal{V}\}$ as input. Here, $\mathcal{F} = \{\Delta_i\}$ denotes the set of triangular faces and $\mathcal{V} = \{v_i\}$ represents the set of vertices.

Overall Architecture. As illustrated in Figure 2, our SchellingFormer begins with a *Hybrid Local Geometric Embedding* module that encodes handcrafted features for each mesh face. These embeddings are processed by a U-Net-style encoder-decoder with skip connections to preserve spatial detail. The encoder comprises five hierarchical stages with progressive downsampling, where each stage contains SchellingFormer layers equipped with our *Laplacian Matrix-Guided Vector Attention* module for structure-aware feature aggregation. Transition Down/Up modules perform resolution changes, while skip connections fuse high-level semantics with local geometry. At the output, each face center receives a predicted saliency score via an MLP. To align with vertex-level Schelling point definitions (Chen et al. 2012), face-level saliency is aggregated to vertices by averaging the scores of adjacent faces. A final normalization is applied across the mesh to produce the final saliency map, from which Schelling points are extracted. In the following, we will introduce our *Hybrid Local Geometric Embedding* module and *Laplacian Matrix-Guided Vector Attention* module in detail.

Hybrid Local Geometric Feature Embedding

Most prevalent Transformer architectures focus on processing pure point clouds as input for modeling long-range spatial relationships and extracting geometric features. However, such coordinate representation fails to explicitly capture important topological and differential geometric cues inherently available in surface meshes (*e.g.*, face connectivity, local continuity and curvature information). This structural deficiency limits the Transformer's ability to reason about surface variation, particularly in tasks like Schelling point detection, which demand a fine-grained understanding of both local shape characteristics and global contextual saliency.

To mitigate this issue, we innovatively augment the input representation with two carefully selected handcrafted geometric descriptors, including the Gaussian curvature and Gaussian curvature difference, derived from surface meshes. These descriptors encode curvature-based surface characteristics that are perceptually and structurally meaningful, serving as a strong local geometric prior for Transformer-based saliency prediction. In the following, we elaborate on the design and motivation behind these two handcrafted descriptors adopted in our approach:

(i) Gaussian Curvature. Gaussian curvature is an intrinsic measure of surface geometry computed as the product of principal curvatures at a point. It characterizes local surface behavior—whether it is convex, concave, saddle-shaped, or flat—and serves as a strong indicator of geometric saliency, as perceptually significant points (e.g., corners, ridges, or junctions) often lie in regions of high curvature variation.

To robustly estimate vertex-wise Gaussian curvature on triangle meshes, we adopt the angle deficit method (Hartig 2021). For a vertex $v_i \in \mathcal{V}$, its curvature is defined as:

$$\mathcal{G}_{v_i} = 2\pi - \sum_j \theta_j, \tag{1}$$

where θ_j denotes the interior angle at v_i in the j-th adjacent face. Intuitively, this measures how much the local surface deviates from being flat. To propagate this signal to the face level, we average the curvature values of the three vertices in each triangular face $\Delta_i \in \mathcal{F}$:

$$\mathcal{G}_{\Delta_i} = \frac{1}{3}\left(\mathcal{G}_{\Delta_{v1}} + \mathcal{G}_{\Delta_{v2}} + \mathcal{G}_{\Delta_{v3}}\right). \tag{2}$$

This face-level Gaussian curvature representation provides a dense and stable local geometric cue that complements positional encoding and enhances the network's sensitivity to perceptually salient regions.

(ii) Gaussian Curvature Difference. To further leverage the intrinsic structure and topological regularity of surface meshes, we compute the Gaussian curvature difference between each face and its local neighborhood to explicitly encode local geometric context. For each face $\Delta_i \in \mathcal{F}$, we define a fixed 9-face patch (as illustrated in Figure 3) consisting of its immediate 1-ring and 2-ring neighboring faces. The curvature difference is then computed as:

$$\mathcal{G}_{\Delta_i}^{\text{diff}} = \left| \mathcal{G}_{\Delta_i} - \frac{1}{9} \sum_j \mathcal{G}_{\Delta_j} \right|, \tag{3}$$

where Δ_j denotes the faces within the defined patch. Compared to vertex-based curvature descriptors, this face-centered formulation provides a more stable and consistent local neighborhood information, as each triangle face has exactly three adjacent faces, whereas vertex valences can vary significantly across the mesh. This regularity mitigates topological inconsistency and improves the reliability of our hybrid local geometric embedding.

Consequently, for each triangular face $\Delta_i \in \mathcal{F}$, we leverage the vector concatenation operation to incorporate the handcrafted local geometric descriptors above, along with the face-centered positions $p_i \in \mathbb{R}^3$, into an unified feature embedding as follows:

$$\mathbf{f}_i = \text{Cat}(\left[p_i,\ \mathcal{G}_{\Delta_i},\ \mathcal{G}_{\Delta_i}^{\text{diff}} \right]) \in \mathbb{R}^5. \tag{4}$$

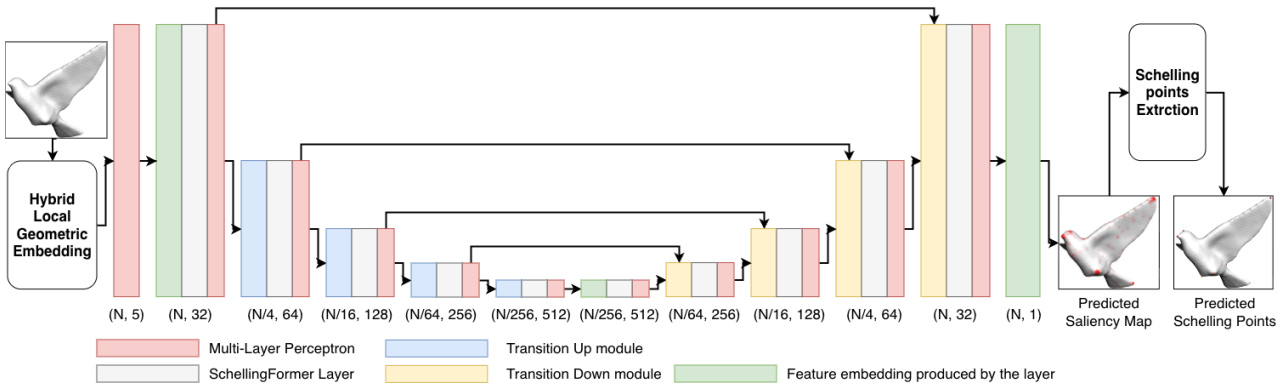

Figure 2: Overview of the proposed framework for Schelling point prediction.

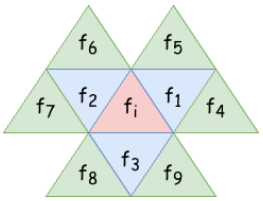

Figure 3: Visualization of the 9-face patch structure.

We note that the centroid coordinates of triangular faces can serve as fundamental spatial anchors, providing SchellingFormer with absolute positional cues over the mesh surface. Unlike relative geometric descriptors, such absolute positional information enables the model to learn high-level spatial layout and distribution patterns across the object. This is particularly useful for distinguishing symmetric but semantically distinct regions (e.g., the head versus the tail of an animal mesh), where curvature-based features alone may be insufficient for disambiguation.

Laplacian Matrix-Guided Vector Attention Layer

Building upon the hybrid local feature embeddings $\{\mathbf{f}_i\}$ introduced above, we feed them into the SchellingFormer network to model long-range dependencies and learn discriminative geometric representations for robust Schelling point prediction. We note that in most existing Transformer-based models, the attention mechanism primarily relies on raw positional encodings while neglecting critical topological cues inherent to surface meshes. This lack of geometric guidance limits the model's ability to propagate information across semantically or structurally meaningful regions, especially for shapes with complex topology. As a result, feature aggregation of Transformer's attention mechanism may become unreliable and misaligned with perceptual saliency.

To address this, we propose a *Laplacian Matrix-Guided Vector Attention* module that incorporates mesh-aware spatial priors into the attention mechanism. Unlike conventional geometry-agnostic self-attention mechanism, our approach explicitly leverages the Laplacian matrix to encode local structural relationships and translates its entries into learnable attention vectors. These attention vectors enable adaptive, structure-aware message passing across the mesh, enhancing both local sensitivity and global contextual reasoning. We find that this design can significantly improve the model's ability to capture intricate geometric structures and form coherent surface-aware representations, both of which are essential for accurate and perceptually aligned Schelling point prediction. In the following, we detail the construction of the Laplacian matrix and the design of the Laplacian matrix-guided vector attention layer:

Laplacian Matrix Establishment. To capture the intrinsic structural relationships within the object surface, we introduce the Laplacian matrix as a spectral geometric prior to guide the Transformer attention, as shown in Figure 4. The Laplacian matrix effectively encodes local connectivity and spatial smoothness, allowing the network to reason about relative structure beyond raw coordinate distances. This is particularly valuable for capturing geometric feature dependencies and enabling geometry-aware message passing.

Specifically, we leverage a set of points $\{p_i\}$ from the centroids of mesh triangles (as described in sub-section above) to construct a localized graph for each point using its K nearest neighbors. We then compute a Laplacian-based affinity matrix $\mathbf{L}$ using a Gaussian kernel applied to the Euclidean distances between each point and its neighbors. The unnormalized affinity weight $\mathbf{L}_{ij}$ between point p_i and its neighbor p_j is defined as:

$$\mathbf{L}_{ij} = \exp\left(-\frac{\|p_i - p_j\|_2^2}{\sigma^2} + \varepsilon\right), \tag{5}$$

where $\|p_i - p_j\|_2$ denotes the Euclidean distance between p_i and p_j; σ is a scaling parameter controlling the kernel width, and ε is a small constant added for numerical stability. To enforce the structural consistency, we explicitly set the self-loop weight $\mathbf{L}_{ii} = 1$. The weights corresponding to the other neighbors are then normalized as:

$$\tilde{\mathbf{L}}_{ij} = \frac{\mathbf{L}_{ij}}{\sum_{k \neq i} \mathbf{L}_{ik} + \delta}, \quad j \neq i, \tag{6}$$

where δ is a small constant to avoid division by zero. This normalization ensures that the weights for each point's neighborhood sum to a consistent scale while preserving the

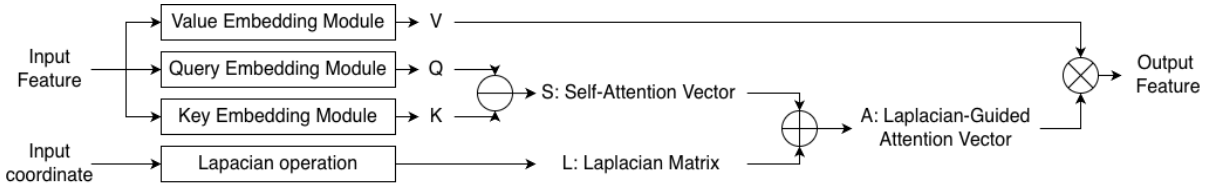

Figure 4: Schellingformer layer.

self-loop. The resulting normalized affinity matrix $\tilde{\mathbf{L}}$ is then used as a spatial relationship-aware prior to guide our Laplacian vector-based attention mechanism.

Laplacian Matrix-Guided Vector Attention. Based on the Laplacian matrix $\tilde{\mathbf{L}}$ introduced above, we propose a Laplacian Matrix-Guided Vector Attention (LMVA) mechanism integrated within the SchellingFormer framework. The key insight of our design is to explicitly leverage spectral geometric priors encoded by $\tilde{\mathbf{L}}$, enabling the attention mechanism to incorporate the structural context for enhanced geometry-aware message passing.

Inspired by the vector self-attention mechanism in Point Transformer, our approach also follows this design to allow attention weights to be adaptively modulated across feature channels rather than uniformly applied. In detail, given the hybrid local geometric features $\mathbf{f}_i$ and $\mathbf{f}_j$ of the face centers p_i and $p_j \in \mathcal{N}_i$ ($\mathcal{N}_i$ denotes the kNN points of p_i) as described in the previous subsection, we first project them into the query $\mathbf{q}_i = \phi_k(\mathbf{f}_i)$ and key $\mathbf{k}_j = \phi_k(\mathbf{f}_j)$ vectors via the respective projection functions, respectively. Then, the preliminary attention input can be computed as:

$$\mathbf{z}_{ij} = \mathbf{q}_i - \mathbf{k}_j + \phi(p_i - p_j), \tag{7}$$

where $\phi(\cdot)$ denotes a learnable positional encoding function that maps the relative offset $p_i - p_j$ into a higher-dimensional position embedding. To further incorporate structure-aware geometric context into the attention mechanism, we concatenate the normalized Laplacian matrix entry $\tilde{\mathbf{L}}_{ij}$ to the preliminary attention input $\mathbf{z}_{i,j}$ along the channel dimension:

$$\tilde{\mathbf{z}}_{ij} = \mathrm{Cat}\left([\mathbf{z}_{ij}, \tilde{\mathbf{L}}_{ij}]\right). \tag{8}$$

This Laplacian-enriched descriptor is then passed through a MLP, followed by a softmax normalization over the neighborhood dimension, to yield vectorized attention weights α_{ij} as follows:

$$\alpha_{ij} = \mathrm{softmax}_j\left(\mathrm{MLP}(\tilde{\mathbf{z}}_{ij})\right). \tag{9}$$

This attention vector is finally used to aggregate the neighborhood value features $\mathbf{v}_j = \phi_v(\mathbf{f}_j)$ in channel-wise manner, producing context-aware representations that capture both local geometry and global structure, which can be formulated as follows:

$$\mathbf{y}_i = \sum_{j \in \mathcal{N}(i)} \alpha_{ij} \odot \mathbf{v}_j, \tag{10}$$

where $\odot$ denotes the Hadamard product operation for element-wise multiplication computation.

Saliency Regression and Schelling Point Detection

After extracting contextual geometric representations with our SchellingFormer, we introduce a saliency regression head that takes the learned geometric features of all face-center points as input to produce a point-wise saliency map over the mesh faces. The predicted saliency scores are subsequently normalized across all face centers to ensure global consistency. Consequently, to identify vertex-level Schelling points, as defined in (Chen et al. 2012), we apply a Schelling point extraction module. First, face-based saliency predictions are converted to vertex-level scores by aggregating the saliency values of each vertex's one-ring adjacent faces. Candidate vertices are then selected based on two criteria: (1) a saliency threshold t, retaining only the top-$t\%$ most salient vertices, and (2) a spatial separation constraint d, which enforces a minimum geodesic distance between selected points via a greedy selection strategy.

Loss Functions

Inspired by the region-aware formulation in (Chen et al. 2022), we adopt a region-aware saliency regression loss for training our Schelling saliency prediction model. Given the predicted saliency map H and the ground-truth saliency map T, the loss is computed by separately considering high-saliency and low-saliency regions. Specifically, we define a scalar threshold τ to distinguish salient from non-salient points. Points with saliency scores above τ are considered in-mask and belong to the set $\mathcal{A}_I$, while points with scores below or equal to τ belong to the out-of-mask set $\mathcal{B}_I$. We then compute the in-mask and out-of-mask losses as:

$$\mathcal{L}_{\text{in}} = \sum_{i \in \mathcal{A}_I} (H_i - T_i)^2, \quad \mathcal{L}_{\text{out}} = \sum_{i \in \mathcal{B}_I} (H_i - T_i)^2 \tag{11}$$

where H_i and T_i denote the predicted and ground-truth saliency scores for point i, respectively. The final region-aware loss is defined as a combination of the two components:

$$\mathcal{L}_{\text{total}} = (1 - \lambda)\mathcal{L}_{\text{in}} + \lambda \mathcal{L}_{\text{out}} \tag{12}$$

where $\lambda \in [0, 1]$ is a coefficient, balancing the relative importance of the out-of-mask region. This weighted loss encourages SchellingFormer to focus more on accurately predicting high-saliency regions, rather than being overwhelmed by the larger low-saliency background. By decoupling the loss contributions from high-saliency and low-saliency regions, the proposed objective yields more stable gradient signals and facilitates saliency-aware learning in this region-sensitive context.

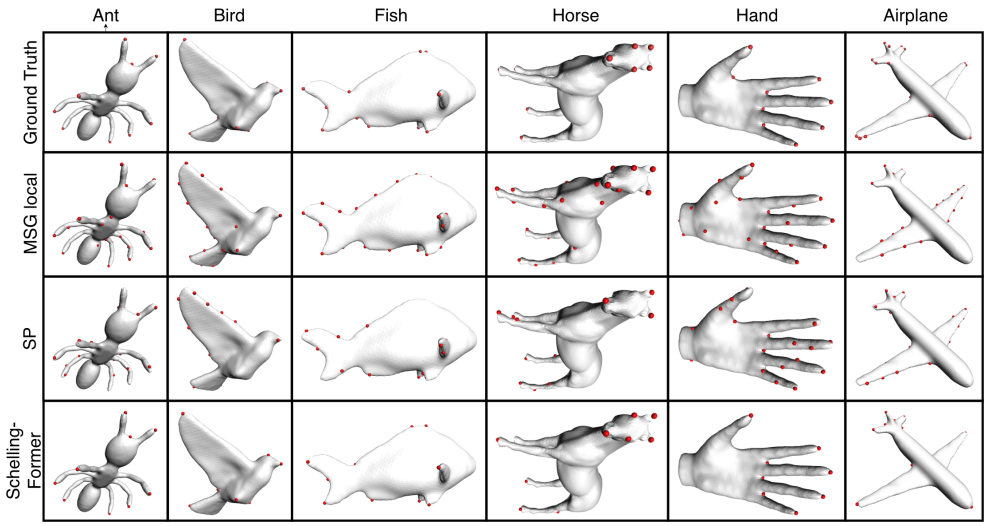

Figure 5: Visual comparison of Schelling point maps. SchellingFormer accurately identifies salient points while significantly reducing false detections.

Experiments

Experimental Setting

Dataset preprocessing. For consistency and fair comparison with existing methods, we use the dataset introduced in (Chen et al. 2012).

Implementation Details. Following (Zhao et al. 2021), we train the model using SGD with a momentum of 0.9, weight decay of 0.0001, an initial learning rate of 0.05, and 80 training epochs. For the region-aware loss, we set the saliency threshold $\tau = 0.015$ and weight $\lambda = 0.1$. The model is implemented in `PyTorch 1.9` and trained on a machine with three NVIDIA GeForce GTX 1080 Ti GPUs (12GB) and an Intel® Core™ i7-7700 CPU (3.60GHz, 8 cores).

Evaluation Metrics. Following (Chen et al. 2022), we report the following detection metrics: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN), and derive F1 score, true positive rate (TPR), false discovery rate (FDR), and mean detection error (MDE) for performance evaluation.

Comparison with Existing Methods

We evaluate all methods on the Schelling point dataset (Chen et al. 2012), using the split from (Chen et al. 2022) (340 train / 60 test) and modifying preprocessing to produce face-level saliency maps. As shown in Table 1, unsupervised baselines (MSG (Lee, Varshney, and Jacobs 2005), SP (Song et al. 2014)) perform poorly, while learning-based models (PointNet++ (Qi et al. 2017), DynGraphCNN (Wang et al. 2019) attain moderately improved results, especially with SPHM integration. CA-SAMNet (Shu et al. 2023) and MMA-POINet (Shu et al. 2024), two of the most recent methods dedicated to POI detection, also demonstrate strong performance. Among existing baselines, MMA-POINet (Shu et al. 2024) achieves the best results, but our SchellingFormer surpasses all methods, achieving the highest F1 (0.6507, +23.7%) and TPR (0.6019, +34.2%), with lower FDR (0.2483) and MDE (0.01598). This performance is driven by two key components: (1) hybrid geometric feature embedding and (2) Laplacian-guided attention for long-range, geometry-aware message passing. Fig. 5 visually confirms these results: SchellingFormer accurately detects most ground-truth Schelling points with fewer false positives, while MSG and SP produce outputs cluttered with false detections.

Ablation Studies and Analysis

Transformer vs. Traditional CNN. We first evaluate the impact of Transformer-based architectures on Schelling point detection. Among CNN-based baselines, DS-Net (built on MeshCNN (Hanocka et al. 2019)) represents a state-of-the-art CNN approach. As shown in Table 3, DS-Net achieves an F1 score of 0.4872, a TPR of 0.4176, an FDR of 0.3709, and an MDE of 0.0201. In contrast, our Transformer-based model, SchellingFormer, achieves substantial improvements across all four metrics. It obtains an F1 score of 0.6507, representing a 33.5% increase over DS-Net, along with a TPR of 0.6019 (a 44.2% improvement), a reduced FDR of 0.2483 (33.1% lower), and an MDE of 0.0160 (20.4% lower). These results represent the superiority of Transformer architectures in capturing long-range geometric dependencies and contextual surface cues essential for accurate Schelling point detection.

Geometric Hybrid Feature Embedding. Then, we evaluate the effect of progressively incorporating handcrafted geometric features into the initial embedding. As shown in the first three rows of Table 2, using only **raw coordinates** provides a strong baseline (F1: 0.5471, TPR: 0.5004), but lacks geometric context, leading to higher FDR in flat or repetitive regions. Adding **Gaussian curvature** yields a substantial improvement in F1 (16.8%) and a significant FDR reduction (28.3%), highlighting its alignment with human visual sensitivity to curvature. Incorporating the **Gaussian curva-**

Method	F1 ↑	TPR ↑	FDR ↓	MDE ↓
MSG (Lee, Varshney, and Jacobs 2005)	0.1934	–	–	–
SP (Song et al. 2014)	0.3232	–	–	–
PointNet++ (Qi et al. 2017)	0.4103	0.3920	0.5089	0.0216
PointNet++*	0.4498	0.4129	0.4217	0.0205
DynGraphCNN (Wang et al. 2019)	0.4347	0.3875	0.4280	0.0209
DynGraphCNN*	0.4699	0.4049	0.3929	0.0204
DS-Net (Chen et al. 2022)	0.4872	0.4176	0.3709	0.0201
CA-SAMNet (Shu et al. 2023)	0.5221	0.4366	0.3514	0.0211
MMA-POINet (Shu et al. 2024)	0.5258	0.4487	0.3633	0.0203
SchellingFormer (Ours)	**0.6507**	**0.6019**	**0.2483**	**0.0160**

Table 1: Quantitative comparison of the proposed method and baselines. Each method is evaluated using F1 score, TPR, FDR, and MDE. Higher values indicate better performance for F1 and TPR, while lower values are preferred for FDR and MDE. Methods marked with * are integrated into the SPHM regression framework proposed by (Chen et al. 2022) for fair comparison.

	Features				Metrics			
Method	Coordinate	Gaussian	GaussDiff	Laplacian	F1 ↑	TPR ↑	FDR ↓	MDE ↓
SchellingFormer	✓				0.54705	0.5004	0.35641	0.01809
SchellingFormer	✓	✓			0.63927	0.58878	0.25572	**0.01584**
SchellingFormer	✓	✓	✓		0.64224	0.5874	**0.24803**	0.01689
SchellingFormer	✓	✓	✓	✓	**0.65068**	**0.60187**	0.24827	0.01598

Table 2: Evaluating the impact of different features and the Laplacian-guided attention mechanism on saliency prediction. Coordinate denotes the point's raw position; Gaussian refers to Gaussian curvature; and GaussDiff refers to the Gaussian Curvature Difference. Feature usage is indicated with checkmarks (✓).

Architecture	F1 ↑	TPR ↑	FDR ↓	MDE ↓
w/ CNN	0.5258	0.4487	0.3633	0.0203
w/ Transformer	**0.6507**	**0.6019**	**0.2483**	**0.0160**

Table 3: Comparison between CNN (DS-NET) and Transformer architectures for Schelling point detection. SchellingFormer consistently outperforms the CNN baseline across all metrics.

ture difference (GaussDiff) further enhances local contrast awareness, providing marginal additional gains. This stepwise enrichment of geometric features consistently improves F1, TPR, and FDR.

Laplacian Matrix-Guided Attention. Finally, we evaluate the contribution of the proposed Laplacian matrix-guided attention module by comparing the performance of SchellingFormer with and without this component under identical feature settings (Rows 3 and 4 in Table 2). Incorporating the Laplacian-guided attention improves the F1 score from 0.6422 to 0.6507 and raises the TPR from 0.5874 to 0.6019, reflecting relative gains of 1.3% and 2.5%, respectively. This highlights the effectiveness of our novel Laplacian matrix-guided vector attention module, which incorporates Laplacian cues as spectral geometric priors to enhance the modeling of long-range relationships. By mapping Laplacian matrix entries to attention vectors, the module enables flexible and geometry-aware message passing, effectively capturing contextual surface cues across the mesh and further help Schelling point detection on mesh.

Conclusions

We have presented a novel Laplacian matrix-guided Geometric Transformer that effectively captures long-range dependencies and discriminative geometric features for robust Schelling point prediction. By incorporating adaptive, geometry-aware message passing, our approach facilitates contextual representation learning that aligns closely with human visual perception. This work bridges the gap between spectral mesh analysis and Transformer-based architectures, offering a unified and powerful framework for 3D shape understanding tasks, including shape matching and saliency detection. Extensive experimental results across multiple benchmarks validate the effectiveness and generalizability of our method, consistently outperforming existing state-of-the-art techniques.

Our method is currently tailored for 2-manifold surface meshes, and it does not explicitly handle the complexities introduced by non-manifold structures. Additionally, it focuses solely on geometric features, without incorporating texture or color information commonly present in real-world 3D models used in applications like film and gaming. In future work, we plan to extend our framework to support irregular geometries and integrate texture cues into the feature embedding process. This integration could further improve the robustness and accuracy of Schelling point detection.

Acknowledgments

This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL), and also supported by MOE AcRF Tier 1 Grant of Singapore (RG12/22).

References

Bai, X.; Luo, Z.; Zhou, L.; Fu, H.; Quan, L.; and Tai, C.-L. 2020. D3feat: Joint learning of dense detection and description of 3d local features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6359–6367.

Castellani, U.; Cristani, M.; Fantoni, S.; and Murino, V. 2008. Sparse points matching by combining 3D mesh saliency with statistical descriptors. In *Computer graphics forum*, volume 27, 643–652. Wiley Online Library.

Chen, G.; Dai, H.; Zhou, T.; Shen, J.; and Shao, L. 2022. Automatic Schelling Point Detection From Meshes. *IEEE Transactions on Visualization and Computer Graphics*, PP: 1–1.

Chen, X.; Saparov, A.; Pang, B.; and Funkhouser, T. 2012. Schelling points on 3D surface meshes. *ACM Transactions on Graphics (TOG)*, 31(4): 1–12.

Dutağacı, H.; Cheung, C.; and Godil, A. 2012. Evaluation of 3D interest point detection techniques via human-generated ground truth. *The Visual Computer*, 28: 901–917.

Fernandez-Labrador, C.; Chhatkuli, A.; Paudel, D. P.; Guerrero, J. J.; Demonceaux, C.; and Gool, L. V. 2020. Unsupervised learning of category-specific symmetric 3d keypoints from point sets. In *European Conference on Computer Vision*, 546–563. Springer.

Gal, R.; and Cohen-Or, D. 2006. Salient geometric features for partial shape matching and similarity. *ACM Trans. Graph.*, 25(1): 130–150.

Golovinskiy, A.; and Funkhouser, T. 2008. Randomized cuts for 3D mesh analysis. *ACM Trans. Graph.*, 27(5).

Hanocka, R.; Hertz, A.; Fish, N.; Giryes, R.; Fleishman, S.; and Cohen-Or, D. 2019. MeshCNN: a network with an edge. *ACM Trans. Graph.*, 38(4).

Hartig, M.-S. 2021. Approximation of Gaussian Curvature by the Angular Defect: An Error Analysis. *Mathematical and Computational Applications*, 26: 15.

He, Y.; Sun, W.; Huang, H.; Liu, J.; Fan, H.; and Sun, J. 2020. PVN3D: A Deep Point-Wise 3D Keypoints Voting Network for 6DoF Pose Estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11629–11638.

Jakab, T.; Tucker, R.; Makadia, A.; Wu, J.; Snavely, N.; and Kanazawa, A. 2021. Keypointdeformer: Unsupervised 3d keypoint discovery for shape control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12783–12792.

Ji, Z.; Liu, L.; Chen, Z.; and Wang, G. 2006. Easy Mesh Cutting. *Comput. Graph. Forum*, 25: 283–291.

Jiang, H.; Dang, Z.; Wei, Z.; Xie, J.; Yang, J.; and Salzmann, M. 2023a. Robust Outlier Rejection for 3D Registration With Variational Bayes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1148–1157.

Jiang, H.; Salzmann, M.; Dang, Z.; Xie, J.; and Yang, J. 2023b. Se (3) diffusion model-based point cloud registration for robust 6d object pose estimation. *Advances in Neural Information Processing Systems*, 36: 21285–21297.

Jiang, H.; Shen, Y.; Xie, J.; Li, J.; Qian, J.; and Yang, J. 2021. Sampling network guided cross-entropy method for unsupervised point cloud registration. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6128–6137.

Kaick, O. V.; Fish, N.; Kleiman, Y.; Asafi, S.; and Cohen-Or, D. 2014. Shape segmentation by approximate convexity analysis. *ACM Transactions on Graphics (TOG)*, 34(1): 1–11.

Katz, S.; Leifman, G.; and Tal, A. 2005. Mesh segmentation using . . . *The Visual Computer*, 21: 649–658.

Katz, S.; and Tal, A. 2003. Hierarchical mesh decomposition using fuzzy clustering and cuts. *ACM Trans. Graph.*, 22(3): 954–961.

Kim, Y.; and Varshney, A. 2006. Saliency-guided Enhancement for Volume Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(5): 925–932.

Kim, Y.; Varshney, A.; Jacobs, D.; and Guimbretiere, F. 2010. Mesh Saliency and Human Eye Fixations. *TAP*, 7.

Lee, C. H.; Varshney, A.; and Jacobs, D. W. 2005. Mesh saliency. In *ACM SIGGRAPH 2005 Papers*, 659–666.

Leifman, G.; Shtrom, E.; and Tal, A. 2016. Surface Regions of Interest for Viewpoint Selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12): 2544–2556.

Li, J.; and Lee, G. 2019. USIP: Unsupervised Stable Interest Point Detection From 3D Point Clouds. 361–370.

Lian, Z.; Godil, A.; Bustos, B.; Daoudi, M.; Hermans, J.; Kawamura, S.; Kurita, Y.; Lavoua, G.; Suetens, P. D.; et al. 2011. Shape retrieval on non-rigid 3D watertight meshes. In *Eurographics workshop on 3d object retrieval (3DOR)*. Citeseer.

Limper, M.; Kuijper, A.; and Fellner, D. W. 2016. Mesh saliency analysis via local curvature entropy. EG '16, 13–16. Goslar, DEU: Eurographics Association.

Liu, R.; and Zhang, H. 2007. Mesh segmentation via spectral embedding and contour analysis. In *Computer Graphics Forum*, volume 26, 385–394. Wiley Online Library.

Mantiuk, R.; Myszkowski, K.; and Pattanaik, S. 2004. Attention Guided MPEG Compression for Computer Animations.

Menzel, N.; and Guthe, M. 2010. Towards perceptual simplification of models with arbitrary materials. In *Computer Graphics Forum*, volume 29, 2261–2270. Wiley Online Library.

Nouri, A.; Charrier, C.; and Lézoray, O. 2015. Multi-scale mesh saliency with local adaptive patches for viewpoint selection. *Signal Processing: Image Communication*, 38: 151–166. Recent Advances in Saliency Models, Applications and Evaluations.

Novatnack, J. 2007. Scale-Dependent 3D Geometric Features. 1 – 8.

Potamias, R. A.; Ploumpis, S.; and Zafeiriou, S. 2022. Neural mesh simplification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18583–18592.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Shilane, P.; and Funkhouser, T. 2007. Distinctive regions of 3D surfaces. *ACM Trans. Graph.*, 26(2): 7–es.

Shu, Z.; Gao, L.; Yi, S.; Wu, F.; Ding, X.; Wan, T.; and Xin, S. 2023. Context-aware 3D points of interest detection via spatial attention mechanism. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6): 1–19.

Shu, Z.; Shen, X.; Xin, S.; Chang, Q.; Feng, J.; Kavan, L.; and Liu, L. 2019. Scribble-based 3D shape segmentation via weakly-supervised learning. *IEEE transactions on visualization and computer graphics*, 26(8): 2671–2682.

Shu, Z.; Yang, S.; Wu, H.; Xin, S.; Pang, C.; Kavan, L.; and Liu, L. 2022a. 3D shape segmentation using soft density peak clustering and semi-supervised learning. *Computer-Aided Design*, 145: 103181.

Shu, Z.; Yang, S.; Xin, S.; Pang, C.; Jin, X.; Kavan, L.; and Liu, L. 2022b. Detecting 3D Points of Interest Using Projective Neural Networks. *IEEE Transactions on Multimedia*, 24: 1637–1650.

Shu, Z.; Yu, J.; Chao, K.; Xin, S.-Q.; and Liu, L. 2024. A Multi-Modal Attention-Based Approach for Points of Interest Detection on 3D Shapes. *IEEE transactions on visualization and computer graphics*, PP.

Sipiran, I.; and Bustos, B. 2011. Harris 3D: A robust extension of the Harris operator for interest point detection on 3D meshes. *The Visual Computer*, 27: 963–976.

Song, R.; Liu, Y.; Martin, R. R.; and Echavarria, K. R. 2018. Local-to-global mesh saliency. *The Visual Computer*, 34(3): 323–336.

Song, R.; Liu, Y.; Martin, R. R.; and Rosin, P. L. 2014. Mesh saliency via spectral processing. *ACM Transactions On Graphics (TOG)*, 33(1): 1–17.

Su, H.; Maji, S.; Kalogerakis, E.; and Learned-Miller, E. 2015. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, 945–953.

Sung, M.; Su, H.; Yu, R.; and Guibas, L. J. 2018. Deep functional dictionaries: Learning consistent semantic structures on 3d models from functions. *Advances in Neural Information Processing Systems*, 31.

Wang, S.; Li, N.; Li, S.; Luo, Z.; Su, Z.; and Qin, H. 2015. Multi-scale mesh saliency based on low-rank and sparse analysis in shape feature space. *Computer Aided Geometric Design*, 35-36: 206–214. Geometric Modeling and Processing 2015.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12.

Wei, G.; Ma, L.; Wang, C.; Desrosiers, C.; and Zhou, Y. 2021. Multi-Task Joint Learning of 3D Keypoint Saliency and Correspondence Estimation. *Computer-Aided Design*, 141: 103105.

Wu, J.; Shen, X.; Zhu, W.; and Liu, L. 2013. Mesh saliency with global rarity. *Graphical Models*, 75(5): 255–264.

Yee, H.; Pattanaik, S.; and Greenberg, D. P. 2001. Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *ACM Trans. Graph.*, 20(1): 39–65.

You, Y.; Liu, W.; Ze, Y.; Li, Y.-L.; Wang, W.-M.; and Lu, C. 2022. UKPGAN: A General Self-Supervised Keypoint Detector. 17021–17030.

Zhao, H.; Jiang, L.; Jia, J.; Torr, P. H.; and Koltun, V. 2021. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16259–16268.

Zhu, X.; Du, D.; Huang, H.; Ma, C.; and Han, X. 2023. 3D Keypoint Estimation Using Implicit Representation Learning. *Computer Graphics Forum*, 42(5): e14917.