2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Manifold Geometry-Based Feature Decoupling for Endoscopic Image Analysis

Yan Wen[1], Haodong Wang[1], Lingyu Chen[2], Wenbo She[1], Dingpei Han[3], Fang Chen[1], Tianqi Huang[1,*]

Abstract— **Endoscopic images suffer from ambiguous semantic boundaries and weak feature discriminability due to acquisition limitations and structural constraints of the anatomical lumen, severely limiting the performance of image analysis models. Mainstream models operate in Euclidean space, which is inherently limited in representing the non-linear geometric characteristics of endoscopic imagery. Manifold space provides a natural advantage in representing such complex structures. Inspired by the manifold hypothesis, this paper models endoscopic image features as separable semantic subspaces on a manifold and proposes a Manifold Geometry-Based Feature Decoupling Module (MANDE). Guided by novel manifold geometric constraints, MANDE adaptively decouples feature maps into multiple semantically independent sub-feature maps, effectively mitigating performance degradation caused by feature space coupling. Furthermore, this paper introduces the Semantic Feature Decoupling-and-Processing (SFDP) framework adopting a divide-and-conquer strategy: utilizing a backbone network for feature extraction, MANDE for decoupling, and a Decouple-Aggregation Head for parallel processing and fusion of sub-features. Extensive experiments demonstrate the framework's adaptability and effectiveness. When integrated with various popular networks, SFDP significantly enhances performance on endoscopic tasks: it reduces RMSE by an average of 14.2% for depth estimation and decreases MAE by 10.5% for segmentation, with only 5.10M additional parameters. Unlike prior works, SFDP uniquely integrates manifold geometry with semantic hierarchical modeling for endoscopic images, providing a novel perspective for surgical robot scene understanding: from holistic features to semantic structures.**

I. INTRODUCTION

Robust visual perception is critical for endoscopic robot precision. In endoluminal surgery, narrow and deformable lumen structures combined with optical interferences create bottlenecks of scale and semantic ambiguity. These arise from the environment's high complexity: the absence of stable geometric references hinders consistent spatial feature extraction, while strong specular reflections and low-texture regions obscure semantic boundaries and reduce discriminability. To address these issues, deep learning-based segmentation and depth estimation have thus become key directions for advancing endoscopic vision.

The U-Net architecture proposed by Ronneberger et al. [1] employs an encoder-decoder framework with symmetrically designed skip connections [2], enabling explicit multi-scale feature extraction. This design effectively alleviated scale ambiguity in medical images and rapidly became a dominant paradigm for subsequent medical image analysis models [3], [4]. A series of U-Net variants further enhanced multi-scale perception, for example by deepening feature extraction layers at each down-sampling stage [4], integrating Transformer blocks to strengthen global context modeling [5], and introducing dual-branch structures to jointly optimize spatial resolution and local-global feature fusion [6]. However, the scarcity of annotated depth data forces many approaches to rely on synthetic rendering [7] or self-supervised strategies. Domain shifts and the absence of reliable depth supervision in self-supervised methods amplify the perceptual bottlenecks of endoluminal scenes and significantly limit the generalization capacity of these models.

The nonlinear geometric features of endoluminal images are difficult to represent effectively in Euclidean space. The manifold hypothesis posits that high-dimensional data are sampled from a low-dimensional manifold structure [8], providing a novel perspective for exploring intrinsic image feature patterns. Recent research has incorporated manifold geometry [9]–[11], leveraging its statistical and geometric properties for feature embedding and constraint modeling aligned with data characteristics. Most studies [12]–[14] focus on high-level semantic tasks, embedding features into task-specific manifold spaces and using geometric metrics to enhance discriminability, improving performance in cross-domain adaptation and hierarchical classification. For pixel-wise dense prediction tasks, [15], [16] map pixel-level features to a manifold space, applying geometric constraints to enforce intra-class compactness. This approach achieves better alignment of class feature distributions and enhances model discriminability in segmentation.

Compared to traditional visual application scenarios, endoluminal environments contain fewer semantic categories, such as lesions versus background and proximal versus distal luminal walls. However, inherent perceptual bottlenecks compel the use of deeper network architectures and multi-scale techniques to extract these highly complex features. Inspired by the manifold perspective [8], we shift our focus away from merely enhancing pixel-level feature encoding and decoding. Instead, from the perspective of semantic component composition, we treat the image feature map as a superposition of multiple semantic sub-features, not as a homogeneous entity [16], [17]. Building upon the manifold hypothesis, this paper proposes the following derived hypothesis: after embedding into the manifold space, pixel-level features from different semantic regions in endoscopic images form separable clusters; each represented by a geometric center, which corresponds to its coordinate on the manifold.

[1]School of Biomedical Engineering, Shanghai Jiao Tong University. [2]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. [3]Department of Thoracic Surgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine.
[*]Corresponding Author. huangtianqi@sjtu.edu.cn

Based on this hypothesis, we introduce manifold modeling to the semantic level of endoluminal images. Specifically, by measuring pixel-level feature differences in the manifold space, we decompose the original feature map into multiple semantically decoupled subspaces. Each subspace is then processed independently by a dedicated task head, and the results are finally fused. This approach effectively mitigates semantic ambiguity and feature coupling caused by endoluminal perceptual bottlenecks.

The main contributions are summarized as follows:

- We propose the Manifold Geometry-Based Feature Decoupling Module (MANDE), which projects input feature into a manifold space to decompose it into semantically independent sub-features based on its intrinsic geometric structure and task-specific objectives, without extra annotations. This enhances representation learning in complex endoluminal scenarios.
- We introduce the Semantic Feature Decoupling-and-Processing (SFDP) framework, a "divide-and-conquer" approach leveraging our MANDE module. SFDP uses MANDE to decouple backbone features into distinct semantic components for parallel processing by specialized heads within a Decouple-Aggregation Head (DAH). The final output is a fusion of all head results.
- We evaluate the proposed SFDP framework by integrating it with various backbones for depth estimation and image segmentation tasks. The experiments demonstrate the framework's generalizability and consistent performance improvement across different networks. Furthermore, we provide an analysis of how different task objectives influence the semantic decoupling effectiveness of the SFDP framework.

II. Methods

A. The Architecture of Network

Based on the derived hypothesis, we propose the MANDE module to achieve semantic feature decoupling of feature maps. To fully exploit the module's capability and validate its performance in end-to-end dense prediction tasks, we construct the SFDP framework (Fig. 1).

During the decoupling process, the attention mechanism within the MANDE module iteratively aligns manifold representations and feature distributions. This process dynamically calibrates the manifold representations under the joint influence of manifold geometric constraints and task-specific guidance to enhance their adaptive representational and discriminative capability for characterizing distinct semantic subspaces. Subsequently, the Decouple Query Layer utilizes these optimized manifold representations as semantic anchors to decompose the original feature map into multiple semantically focused sub-feature maps, achieving an unstructured semantic separation.

To ensure both discriminative power and structural consistency of the disentangled features, we introduce two dedicated constraints: the Semantic Diversity Constraint and the Hierarchical Constraint. Each component will be introduced in the following subsections.

B. Manifold Geometry-Based Feature Decoupling Module

Given a feature map $F_0 \in \mathbb{R}^{H\times W\times C}$ from a traditional backbone network, where H and W denote its height and width, and C is the channel dimension. The MANDE module decouples F_0 into K semantically independent sub-feature maps, forming the set $\{F_i\}_{i=1}^{K}$, where each $F_i \in \mathbb{R}^{H\times W\times C}$ and K is a pre-defined hyperparameter representing number of decoupled components.

The structure of the MANDE module is illustrated in Fig. 2. The process begins with a preprocessing stage, where the original feature map F_0 is downsampled to a lower resolution to form a coarse-grained feature sequence. This sequence is used to initialize as K semantic manifolds $\left\{\mathcal{M}_i^{(0)}\right\}_{i=1}^{K}$. The Semantic Manifold Alignment Block (SMAB), which incorporates hybrid attention, iteratively refines the alignment between these manifolds and the underlying semantic features. Upon convergence of the manifolds $\mathcal{M}_i$, the Decouple Query Layer integrates their semantic information with the original features F_0 to produce high-resolution semantically decoupled feature maps.

Preprocessing. Inspired by [12], we convert the feature map F_0 into a low-resolution feature sequence for computational efficiency and map to the manifold space to obtain the global semantic manifold $\mathcal{M}_G$ and the initialized local semantic manifolds $\left\{\mathcal{M}_i^{(0)} \in \mathbb{R}^C \middle| i = 1, 2, \ldots, K\right\}$. Specifically, F_0 is downsampled via a convolutional layer with kernel size of $p \times p$ and stride p , yielding a low-resolution feature map $F_p \in \mathbb{R}^{h\times w\times C}$, where $h = H/p$ and $w = W/p$. This feature map F_p is then reshaped into a semantic sequence $S^0 \in \mathbb{R}^{hw\times C}$. To initialize K independent semantic manifolds, S^0 is replicated K times through K separate linear layers and non-linearly projected into the manifold space by a shared AvgPool-Conv-Activation projector. Origin S^0 becomes the global semantic manifold $\mathcal{M}_G$, while the K replicates form the initial set of local semantic manifolds $\left\{\mathcal{M}_i^{(0)}\right\}_{i=1}^{K}$.

Semantic Manifold Alignment Block. The Semantic Manifold Alignment Block (SMAB) iteratively refines K semantic manifolds, driving them toward convergent and geometrically distinct structures via a Hybrid Attention mechanism. Let $\mathcal{M}^{(t)} = \left\{\mathcal{M}_i^{(t)}\right\}_{i=1}^{K}$ denote the set of manifold representations after the t-th iteration, where $t = 0$ corresponds to the initialized state. After optimization through N cascaded SMABs, the final converged manifold set is denoted as $\mathcal{M}^{(N)} = \{\mathcal{M}_i\}_{i=1}^{K}$. The first SMAB processes the input manifolds $\mathcal{M}^{(0)}$ and the semantic sequence $S^{(0)}$, producing updated outputs $\mathcal{M}^{(1)}$ and $S^{(1)}$. The operation of the t-th SMAB is formally defined as follows:

$$\mathcal{M}^{(t)}, S^{(t)} = \mathrm{SMAB}_t\left(\mathcal{M}^{(t-1)}, S^{(t-1)}\right) \tag{1}$$

Within each block, the operations for each manifold $\mathcal{M}_i^{(t)} \in \mathcal{M}^{(t)}$ and its corresponding sequence $S_i^{(t)} \in S^{(t)}$ are defined as follows:

$$\begin{aligned} H_i^{(t)}, S_i^{(t)} &= \mathrm{HybridAttn}\left(\mathcal{M}_i^{(t-1)}, S_i^{(t-1)}\right) \\ \mathcal{M}_i^{(t)} &= \mathrm{MLP}(\mathrm{LayerNorm}(H_i^{(t)})) + H_i^{(t)} \end{aligned} \tag{2}$$

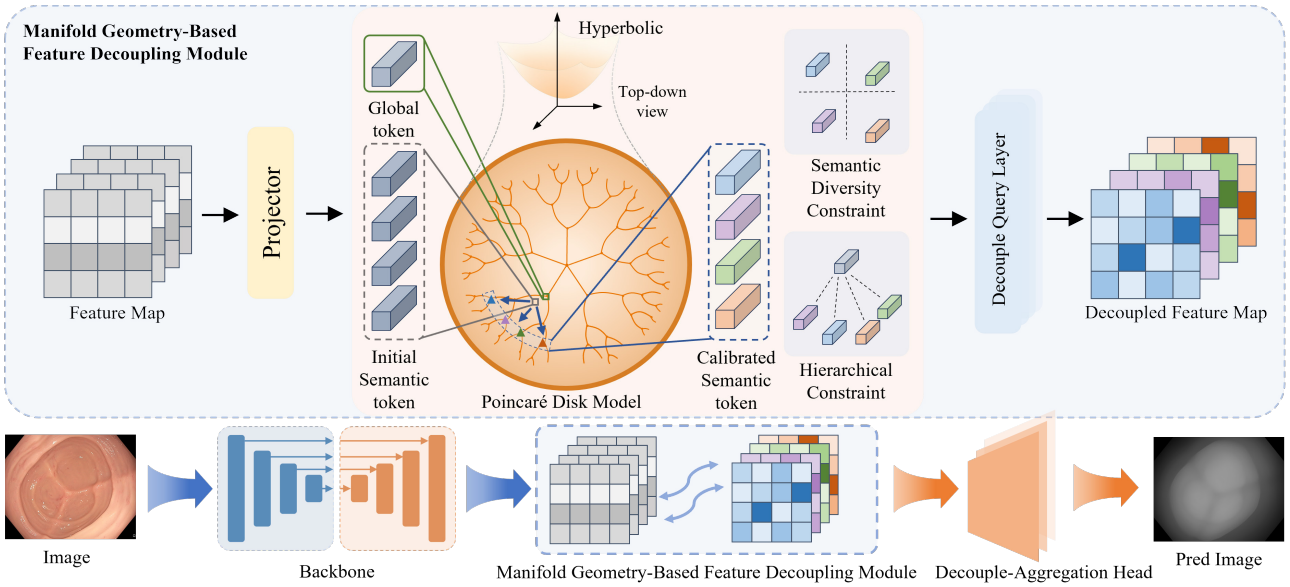

Fig. 1. **Overview of the SFDP framework.** The input image is first processed by a backbone network (e.g., U-Net or Transformer variants) to extract hierarchical feature maps. These features are then passed to the MANDE module, which non-linearly projects them into a manifold space and iteratively refines the manifold representations to align with distinct semantic regions. Finally, the Decouple-Aggregation Head processes the semantically decoupled features using parallel task-specific heads and aggregates their outputs to produce the final prediction.

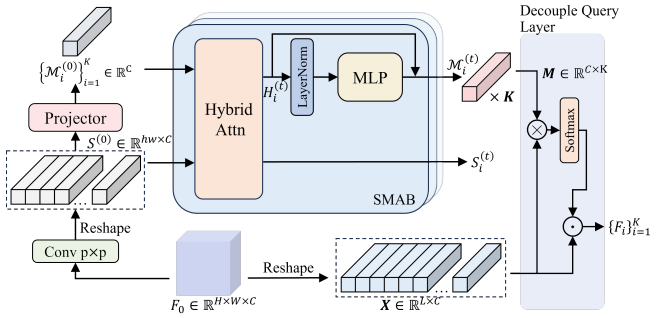

Fig. 2. Manifold Geometry-Based Feature Decoupling Module.

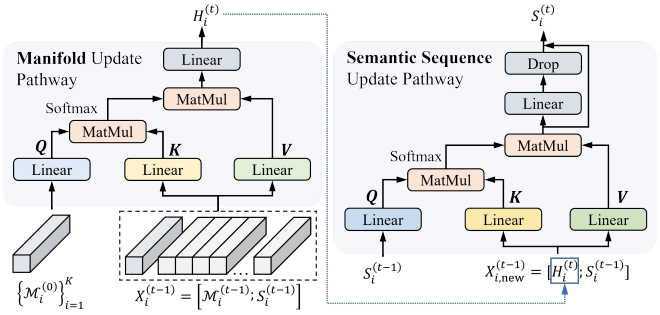

Fig. 3. Hybrid Attention Module.

where HybridAttn denotes Hybrid Attention Module (see Fig. 3 for details). MLP and LayerNorm refer to the Multi-Layer Perceptron and the layer normalization layer, respectively. $H_i^{(t)}$ represents the hidden representation produced by the Hybrid Attention Module.

Hybrid Attention Module. This module facilitates mutual optimization between semantic manifolds $\mathcal{M}_i^{(t)}$ and semantic sequence $S_i^{(t)}$ through an alternating attention mechanism, which consists of two sequential pathways: 1) Manifold Update Pathway: utilizing $\mathcal{M}_i^{(t-1)}$ as the query, it dynamically aggregates critical semantic features from $S_i^{(t-1)}$, generating a more discriminative hidden representation $H_i^{(t)}$ for subsequent update to $\mathcal{M}_i^{(t)}$. 2) Semantic Sequence Update Pathway: using $S_i^{(t-1)}$ as the query and incorporates the updated information from $H_i^{(t)}$ to generate the refined sequence $S_i^{(t)}$, This process facilitates noise reduction while preserving structural details. The two pathways work collaboratively to enhance the discriminability of the manifolds while suppressing semantic interference, thereby providing a stable foundation for the semantic decoupling optimization. Formally, for given input manifold $\mathcal{M}_i^{(t-1)} \in \mathbb{R}^C$ and semantic sequence $S_i^{(t-1)} \in \mathbb{R}^{hw \times C}$, the operations of the two pathways are defined as follows.

In the Manifold Update Pathway, the manifold and the semantic sequence are concatenated to form $X_i^{(t-1)} = [\mathcal{M}_i^{(t-1)}; S_i^{(t-1)}] \in \mathbb{R}^{(hw+1)\times C}$, yielding a complete spatial-semantic description that enables bidirectional alignment and joint optimization between manifold representations and pixel-level features. The process update is formulated as:

$$
\begin{aligned}
&Q_m = W_q^m \mathcal{M}_i^{(t-1)}, K_m = W_k^m X_i^{(t-1)}, V_m = W_v^m X_i^{(t-1)} \\
&A_m = \text{softmax}\left(\frac{Q_m K_m^T}{\sqrt{d}}\right), H_i^{(t)} = W_{\text{proj}}^m (A_m \cdot V_m)
\end{aligned} \tag{3}
$$

where $W_q^m, W_k^m, W_v^m, W_{\text{proj}}^m \in \mathbb{R}^{C\times C}$ are learnable weight matrices, and $d = \frac{C}{\text{head}}$ is the scaling factor, with head denoting the number of heads in the multi-head attention mechanism. $H_i^{(t)}$ denotes the hidden representation.

Similarly, in the Semantic Sequence Update Pathway, $H_i^{(t)}$ and the semantic sequence are concatenated to form $X_{i,\text{new}}^{(t-1)} = [H_i^{(t)}; S_i^{(t-1)}] \in \mathbb{R}^{(hw+1)\times C}$. The update follows:

$$Q_s = W_q^s S_i^{(t-1)},\ K_s = W_k^s X_{i,\text{new}}^{(t-1)},\ V_s = W_v^s X_{i,\text{new}}^{(t-1)}$$
$$A_s = \text{softmax}\left(\frac{Q_s K_s^T}{\sqrt{d}}\right), S_i^{(t)} = W_{\text{proj}}^s (A_s \cdot V_s) + S_i^{(t-1)} \quad (4)$$

where $W_q^s, W_k^s, W_v^s, W_{\text{proj}}^s \in \mathbb{R}^{C\times C}$ are learnable weight matrices. The residual connection in the update for $S_i^{(t)}$ ensures the preservation of original semantic information.

Decouple Query Layer. The Decouple Query Layer generates K semantically decoupled feature maps $\{F_i\}_{i=1}^K$ by leveraging the original high-resolution feature map $F_0 \in \mathbb{R}^{H\times W\times C}$ and the converged manifold representations $\{\mathcal{M}_i\}_{i=1}^K$ from the SMAB, where each $\mathcal{M}_i \in \mathbb{R}^C$ and $F_i \in \mathbb{R}^{H\times W\times C}$. Specifically, the layer operates by computing the similarity between each $\mathcal{M}_i$ and all pixel-wise features in F_0. This computation produces a Decouple Attention Map $A_i \in \mathbb{R}^{H\times W}$, which quantifies the responsiveness of each pixel in F_0 to the semantic meaning represented by $\mathcal{M}_i$. Based on this, semantically relevant features are selectively extracted from F_0, achieving semantic decoupling.

First, the feature map F_0 is reshaped into a matrix $\boldsymbol{X} \in \mathbb{R}^{L\times C}$, where $L = H \times W$. The K manifolds are concatenated to form a matrix $\boldsymbol{M} = [\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_K] \in \mathbb{R}^{C\times K}$. The affinity between $\boldsymbol{X}$ and $\boldsymbol{M}$ is then computed as:

$$\boldsymbol{A} = \text{softmax}\left(\boldsymbol{X}\boldsymbol{M}\right) \in \mathbb{R}^{L\times K} \quad (5)$$

where the softmax operation is applied along the dimension of size K for normalization. Each vector $\boldsymbol{A}_{:,i} \in \mathbb{R}^L$ represents the attention weight vector for the i-th semantic category across all spatial locations, indicating each pixel's responsiveness to $\mathcal{M}_i$ in the original feature F_0.

Subsequently, reshape $\boldsymbol{A}_{:,i}$ into a Decouple Attention Map $A_i \in \mathbb{R}^{H\times W}$. The final decoupled feature map F_i is then obtained via element-wise multiplication:

$$F_i = A_i \odot F_0 \quad (6)$$

where $\odot$ denotes the element-wise multiplication operation. The resulting set of feature maps $\{F_i\}_{i=1}^K$ retains the full spatial resolution, with each F_i focusing on features corresponding to the i-th semantic subspace.

C. Manifold Constraint

To guide the MANDE module towards effective feature decoupling, this section introduces constraints on the manifold representations $\{\mathcal{M}_i\}_{i=1}^K$ and $\mathcal{M}_G$: 1) **Semantic Diversity Constraint:** minimize the cosine similarity between the manifolds $\{\mathcal{M}_i\}_{i=1}^K$ to ensure clear separation and discriminative boundaries among semantic subspaces; 2) **Hierarchical Constraint:** by leveraging hyperbolic space, specifically the Poincaré ball model, we establish a hierarchical constraint between the global manifold $\mathcal{M}_G$ and the local manifolds $\{\mathcal{M}_i\}_{i=1}^K$ using distance metrics defined in the Poincaré model.

Semantic Diversity Constraint. To ensure the decoupled sub-feature maps $\{F_i\}_{i=1}^K$ capture distinct semantic information, we enforce diversity directly on the manifold representations that generate them. Given the K manifolds $\{\mathcal{M}_i\}_{i=1}^K$ obtained from the SMAB, we apply manifold orthogonality constraint, maximizing the angular separation between the manifolds $\mathcal{M}_i$ to minimize semantic redundancy. Consequently, each feature map F_i exhibits strong responsiveness to a unique semantic dimension with maximal uniqueness.

$$L_{\text{sd}} = \frac{1}{K}\sum_{i=1}^{K}\sum_{j\neq i}|\boldsymbol{m}_i^\top \boldsymbol{m}_j| \quad (7)$$

where $\boldsymbol{m}_i = \mathcal{M}_i/\|\mathcal{M}_i\|_2 \in \mathbb{R}^C$ denotes the L2-normalized representation of $\mathcal{M}_i$.

Hierarchical Constraint. To model the inherent hierarchical relationships within the semantic manifolds, we propose two hierarchical constraints: 1) Global-Local: between the global semantic manifold $\mathcal{M}_G$ and each local semantic manifold $\mathcal{M}_i$; 2) Local-Local: among the K local semantic manifolds $\{\mathcal{M}_i\}_{i=1}^K$, which theoretically belong to similar hierarchical levels. Inspired by [13], [18], [19], we introduce hyperbolic space to measure these hierarchical constraints. Hyperbolic space $\mathbb{H}^n$ is an n-dimensional Riemannian manifold with constant negative sectional curvature $-c$. For practical implementation in neural networks, we adopt the Poincaré ball model, to facilitate gradient-based optimization in neural networks. The Poincaré ball is defined as an n-dimensional open ball of radius $1/\sqrt{c}$: $\mathcal{B}^n = \{\mathbf{x} \in \mathbb{R}^n : c\|\mathbf{x}\| < 1\}$, equipped with a conformal metric tensor $g_{\mathbf{x}} = (\frac{2}{1-c\|\mathbf{x}\|^2})^2 g^E$, where g^E denotes the Euclidean metric tensor and $\|\cdot\|$ is the Euclidean norm. The corresponding Riemannian manifold is $(\mathcal{B}^n, g_x)$. Then, the geodesic distance between two points $\mathbf{u}, \mathbf{v} \in \mathcal{B}^n$ is given by:

$$d_{\mathcal{B}}(\mathbf{u},\mathbf{v}) = \frac{1}{\sqrt{c}}\cosh^{-1}\left(1 + 2c\frac{\|\mathbf{u}-\mathbf{v}\|^2}{(1-c\|\mathbf{u}\|^2)(1-c\|\mathbf{v}\|^2)}\right) \quad (8)$$

As can be observed from the formula (8), the geometry of the Poincaré ball exhibits two key regimes: near the origin, where node norms are small, the distance behaves approximately as Euclidean space. In contrast, as nodes approach the boundary, the distance between them grows exponentially. This property is ideal for modeling tree-like structures: the exponential growth characteristic of $d_{\mathcal{B}}(\mathbf{u},\mathbf{v})$ amplifies the differences between leaf nodes on distinct branches. Consequently, nodes with smaller norms are typically embedded closer to the origin and represent higher-level abstract concepts, while those with larger norms are closer the boundary and represent more specific, leaf-level information [11], [14], [18]. In our framework, the global semantic manifold $\mathcal{M}_G$ encapsulates more abstract semantic information than the local semantic manifolds $\{\mathcal{M}_i\}_{i=1}^K$, and is thus treated as a higher-level node. To impose hierarchical constraints, we first project $\mathcal{M}_G$ and $\{\mathcal{M}_i\}_{i=1}^K$ into the Poincaré ball model using the following formula [18]:

$$\text{proj}(\mathbf{x}) = \begin{cases} \mathbf{x}\cdot\dfrac{R}{\|\mathbf{x}\|}, & if \|\mathbf{x}\| > R, R = \dfrac{1}{\sqrt{c}} \\ \mathbf{x}, & otherwise \end{cases} \quad (9)$$

Based on this projection, the semantic manifolds are mapped into the Poincaré ball model, yielding a parent node $\boldsymbol{p} = \text{proj}(\mathcal{M}_G)$ and a set of child nodes

$\{\boldsymbol{c}_i = \mathrm{proj}\left(\mathcal{M}_i\right)\}_{i=1}^{K}$. These embedded representations are constrained by the following hierarchical relationships:

Global-Local Constraint: The hierarchical relationship between the high-level parent node $\boldsymbol{p}$ and the child nodes $\{\boldsymbol{c}_i\}_{i=1}^{K}$ is enforced through three conditions within the Poincaré ball: The **Norm Constraint** ensures children are embedded closer to the boundary via $\|\boldsymbol{c}_i\| > \|\boldsymbol{p}\| + \epsilon/c$; The **Direction Constraint** maintains semantic consistency by enforcing a high cosine similarity $\frac{\boldsymbol{c}_i \cdot \boldsymbol{p}}{\|\boldsymbol{c}_i\|\|\boldsymbol{p}\|} \geq \eta$; Finally, the **Distance Constraint** regularizes the geodesic distance $d(\boldsymbol{p}, \boldsymbol{c}_i)$ to prevent excessively large separation. These constraints can be formally expressed as formula (10), where $[\cdot]_+$ denotes function $x = \max(0, x)$.

$$\begin{aligned} L_{\text{gtl}} = \lambda_1 \sum_{i=1}^{K} \left[\|\boldsymbol{p}\| - \|\boldsymbol{c}_i\| + \frac{\epsilon}{\sqrt{c}} \right]_+ \\ + \lambda_2 \sum_{i=1}^{K} \left[\eta - \frac{\boldsymbol{p} \cdot \boldsymbol{c}_i}{\|\boldsymbol{p}\|\|\boldsymbol{c}_i\|} \right]_+ + \lambda_3 \sqrt{\frac{1}{K} \sum_{i=1}^{K} d(\boldsymbol{p}, \boldsymbol{c}_i)^2} \end{aligned} \tag{10}$$

Local-Local Constraint: To further promote feature separation and mitigate semantic mixing among the K child nodes $\{c_i\}_{i=1}^{K}$, we establish a repulsive constraint in hyperbolic space. This constraint acts as an auxiliary to the Semantic Diversity Constraint, explicitly enforcing a hierarchical dispersion of the child nodes to enhance the discriminability of the decoupled subspaces. This constraint is expressed as follows, where $\tau > 0$ is a temperature parameter, we set $\tau = 0.2$ in this work.

$$L_{\text{ltl}} = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j=1, j \neq i}^{K} \exp\left(-\frac{d_c(\boldsymbol{c}_i, \boldsymbol{c}_j)}{\tau} \right) \tag{11}$$

Combining the global-local constraint L_{gtl} and the local-local constraint L_{ltl}, the overall Hierarchical Constraint is defined as follows: $L_{\text{hier}} = L_{\text{gtl}} + L_{\text{ltl}}$.

In summary, the Manifold Constraint is obtained as:

$$L_{\text{m}} = L_{\text{sd}} + L_{\text{hier}} \tag{12}$$

D. Decouple-Aggregation Head

The Decouple-Aggregation Head (DAH) bridges feature decoupling and downstream pixel-level feature processing tasks. As a lightweight module, DAH processes each decoupled sub-feature independently using dedicated sub-networks, then fuses the outputs. These sub-networks can be modified in future work for specific needs.

Given the set of K decoupled sub-feature maps $\{F_i\}_{i=1}^{K}$ (where each $F_i \in \mathbb{R}^{H \times W \times C}$) derived from the original feature map $F_0 \in \mathbb{R}^{H \times W \times C}$, the DAH processes each sub-feature map F_i with an independent task-specific network. The results from these parallel networks are then dynamically fused using a set of learnable weights. This design preserves the discriminative properties of the decoupled features while ensuring strong adaptability to the downstream task.

In implementation, the DAH consists of K parallel task networks. Each sub-network is a lightweight convolutional module dedicated to processing features from its corresponding semantic subspace. For the i-th decoupled feature map F_i, its sub-network generates a task prediction map D_i through the following structure:

$$D_i = \mathrm{Conv}_{1\times1}\left(\mathrm{ELU}\left(\mathrm{Conv}_{3\times3}\left(F_i\right)\right)\right) \in \mathbb{R}^{H \times W} \tag{13}$$

where $\mathrm{Conv}_{k\times k}$ denotes a $k \times k$ convolutional layer. The two convolutional layers sequentially compress the channel dimension of the feature maps from C to $C/2$ and then to a single-channel output. $\mathrm{ELU}(\cdot)$ represents the ELU activation function. All sub-networks process the set $\{F_i\}_{i=1}^{K}$ in parallel to produce the corresponding set of task predictions $\{D_i\}_{i=1}^{K}$. These predictions are then adaptively fused using a set of learnable weights $\boldsymbol{w} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_K]^T \in \mathbb{R}^K$. The final aggregated prediction is obtained through the Sigmoid activation function $\sigma(\cdot)$:

$$Y = \sigma\left(\sum_{i=1}^{K} \boldsymbol{w}_i \cdot D_i\right) \tag{14}$$

III. Experiments

To demonstrate the effectiveness of our proposed feature decoupling module, we integrate several established baseline models into the SFDP Framework and conduct comparative experiments on two pixel-level tasks: depth estimation and image segmentation. The results validate the efficacy and generalizability of our approach. Furthermore, ablation studies are performed to compare the impacts of decoupling number and manifold constraints on model performance.

A. Model Configuration

Baseline. For depth estimation, we selected models prevalent in autonomous driving and endoscopic imaging for comparative experiments: DepthNet [20]–[22], SwinDepthMIM [23], [24] (a Swin-Transformer-based model pre-trained with masked image modeling), and the multi-task model SwinMTL [24]. For image segmentation, we chose leading models in polyp segmentation: PraNet [25], LSSNet [4]. We also include VM-UNet [3], a recent architecture that integrates Vision Mamba and U-Net.

Framework-Based Optimization. To ensure a fair and consistent evaluation, we integrated the baseline models (DepthNet, SwinMTL, LSSNet, and VM-UNet) into the SFDP framework with uniform modifications. The original prediction heads of these models were removed to repurpose them as feature extractors within our SFDP framework's backbone. The generated feature maps are then processed by the standard SFDP workflow: semantic feature decoupling is performed by the MANDE module, and the final task-specific prediction is generated by the Decouple-Aggregation Head (DAH). All original models utilized parameters reported in their papers, with the backbone output channel dimension uniformly set to $C = 128$ and the number of decoupled semantic components set to $K = 4$.

B. Experimental Setup

Datasets. For depth estimation, we use the C3VD colonoscopy dataset[1] [26]. All images are resized to 256×256. The ground truth depth maps, provided as 16-bit grayscale images with linear scaling, are rescaled to $[0, 10]$.

To evaluate generalization, we use multiple polyp segmentation datasets: EndoScene [27] (912 pairs), Kvasir-SEG [28] (100 randomly selected pairs), CVC-ColonDB [29] (380 pairs), and ETIS [30] (196 pairs). Among these, EndoScene was used exclusively for training and validation according to its default split, while all other segmentation datasets were used for testing. For segmentation, images are resized to 352×352, with label scaled to $[0, 1]$.

Task-Supervised loss. We employ standard task-specific loss functions. For depth estimation, the loss combines the scale-invariant error loss [31] and gradient matching error: $L_{\text{depth}} = L_{SILog} + L_{\text{grad}}$. For image segmentation, we used the structural loss [25]: $L_{\text{seg}} = L_{wBCE} + L_{wIoU}$, combining weighted Binary Cross-Entropy and weighted Intersection-over-Union. The overall training objective is defined as:

$$Loss = L_{\text{sup}} + \alpha_{\text{m}} L_{\text{m}} \tag{15}$$

where L_{sup} denotes the task supervision loss (i.e., L_{depth} for depth estimation or L_{seg} for image segmentation), and L_{m} is the Manifold Constraint, with $\alpha_{\text{m}} = 0.1$.

Implementation Details. All experiments were conducted on a workstation with Intel(R) Core(TM) i7-14700KF CPU, NVIDIA GeForce RTX 4070 Ti SUPER GPU, and 32 GB RAM. The models were trained for 50 epochs with batch size 4, using the AdamW optimizer [32] with initial learning rate of 1e-4. The learning rate was scheduled to decay by a factor of 0.5 every 5 epochs via StepLR scheduler. The weight decay was set to 1e-2 for depth estimation models and VM-UNet, and 1e-4 for other image segmentation models.

C. Effectiveness

Depth Estimation. We evaluate the depth estimation performance of DepthNet and SwinMTL, both in their original forms and integrated with the SFDP framework (denoted as Model_M). The quantitative results in Table I, demonstrate that the SFDP consistently yields statistically significant performance improvements across both architectures. Notably, SFDP reduces the Root Mean Square Error (RMSE) by 11.28% for DepthNet and by 17.09% for SwinMTL. Beyond these, Fig. 4 illustrates that SFDP significantly improves edge perception compared to original networks, particularly in addressing depth boundary regression errors caused by low-texture regions in endoscopic environments.

Image Segmentation. To comprehensively assess SFDP's capability in improving both accuracy and cross-domain generalization, we train models on EndoScene dataset and test on three unseen datasets: Kvasir-SEG, CVC-ColonDB, and ETIS datasets. As presented in Table II, SFDP-enhanced LSSNet and VM-UNet (denoted as Model_M) exhibit consistent performance gains across diverse test domains, with the majority of improvements achieving statistically significance. Specifically, SFDP achieves an average MAE reduction of 14.65% for LSSNet and 6.34% for VM-UNet. Qualitative results in Fig. 5 further demonstrate that the SFDP helps mitigate common failure modes in endoscopic segmentation, such as misjudgments caused by low texture and uneven illumination, leading to more robust and accurate predictions.

[1]We used data from four C3VD sub-datasets: cecum_t1_b (cecum), desc_t4_a (descending colon), sigmoid_t1_a (sigmoid colon), and trans_t1_b (transcending colon).

TABLE I

QUANTITATIVE RESULTS OF DEPTH ESTIMATION EXPERIMENTS.

Model	AbsRel ↓	SqRel ↓	RMSE ↓	RMSE log ↓
SwinDepthMIM	0.0255	0.0039	0.1030	0.0470
DepthNet	0.0212	0.0032	0.1028	0.0350
DepthNet_M	**0.0195***	**0.0026***	**0.0912***	**0.0321***
SwinMTL	0.0270	0.0041	0.1012	0.0496
SwinMTL_M	**0.0222***	**0.0030***	**0.0839***	**0.0425***

† Model_M represents the model integrated with SFDP. Statistical significance (p-value < 0.05) is denoted with *(Model_M vs Model).

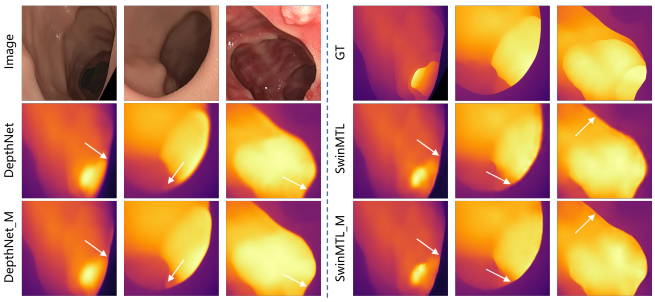

Fig. 4. Qualitative results of depth estimation experiments.

Semantic Decoupling. To analyze how task objectives influence the decoupling behavior, we train the DepthNet_M on distinct tasks: depth estimation using the C3VD dataset and image segmentation using the EndoScene dataset, respectively. Fig. 6 visualizes the Decouple Attention Maps generated by the Decouple Query Layer, which quantify the semantic proximity between feature vectors of each pixel in the original feature map and each learned semantic manifold. These attention weights directly determine how much each pixel contributes to the corresponding decoupled sub-feature map, thereby revealing which semantic regions each manifold captures. The Decouple Attention Maps results clearly show that the learned manifolds successfully focus on distinct and semantically meaningful image regions, validating that MANDE effectively decomposes the feature space into interpretable semantic components.

Furthermore, the task objectives fundamentally guide the semantic decoupling patterns. For depth estimation, the decoupling process focuses on discerning patterns in depth value distribution, effectively learning semantics related to spatial geometry. For instance, the manifolds learn to distinguish between the proximal rectal wall, mid-range structures, and distal luminal areas. In contrast, segmentation task objectives drive the formation of semantically decoupled components that capture categorical distinctions like background, contours, and foreground polyp tissues. This discrepancy reflects the fundamental guidance of task objectives on decoupled feature representation: depth estimation drives the decoupling along geometric and spatial properties, while segmentation produces decoupling that emphasizes boundaries between anatomical structures and tissue types.

TABLE II

QUANTITATIVE RESULTS OF IMAGE SEGMENTATION EXPERIMENTS. ALL DATASETS WERE USED SOLELY FOR TESTING.

Model	KvasirSeg			CVC-ColonDB			ETIS		
	MAE↓	Dice↑	IoU↑	MAE↓	Dice↑	IoU↑	MAE↓	Dice↑	IoU↑
PraNet	0.0714	0.7859	0.7022	0.0377	0.7884	0.7274	0.0198	0.6735	0.5937
LSSNet	0.0519	0.8375	0.7624	0.0369	0.8001	0.7319	0.0231	0.7146	0.6290
LSSNet_M	**0.0435**	**0.8580***	**0.7845***	**0.0324***	**0.8200***	**0.7537***	**0.0195**	**0.7565***	**0.6687***
VM-UNet	0.0506	0.8343	0.7599	0.0383	0.7957	0.7319	0.0232	0.6916	0.5962
VM-Unet_M	**0.0490**	**0.8416**	**0.7691**	**0.0380**	**0.8003**	**0.7343**	**0.0197***	**0.7091***	**0.6190***

† Model_M represents the model integrated with SFDP. Statistical significance (p-value < 0.05) is denoted with *(Model_M vs Model).

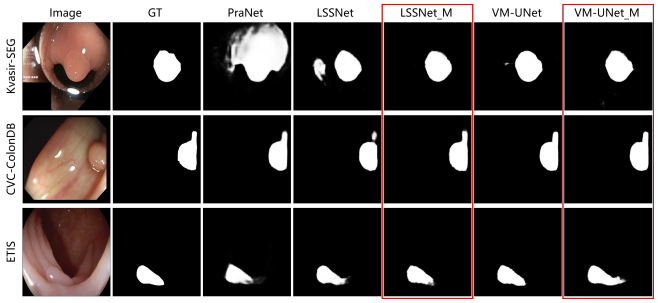

Fig. 5. Qualitative results of image segmentation experiments.

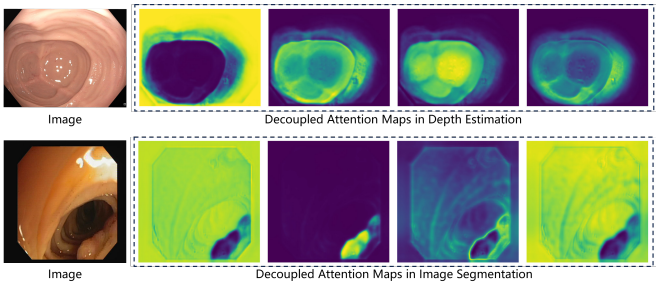

Fig. 6. Decoupled Attention Maps for Different Tasks: Depth Estimation (Top) and Image Segmentation (Bottom).

D. Ablation Studies

We conducted ablation experiments to evaluate key design choices using DepthNet_M, an improved version of the U-Net-based DepthNet, for depth estimation. We investigate the impact of the number of decoupled sub-features and the presence of manifold constraints.

Decoupled Sub-feature Count. The hyperparameter K determines the number of semantic components extracted by the MANDE module and the number of parallel task heads in the DAH. This value represents a critical trade-off. With fixed feature-map dimensions, an insufficiently small K limits the diversity of decoupled semantics, yielding only marginal improvements. Conversely, an excessively large K increases computational complexity, amplifies the challenge of achieving effective decoupling, and introduces redundant components, which can dilute the salience of meaningful features. As shown in Table III, DepthNet_M achieves optimal performance on the depth estimation task at $K = 4$. The results demonstrate that performance does not increase monotonically with K. Instead, exceeding a critical threshold leads to a mismatch between model capacity and constraint complexity, resulting in reduced training stability and significant performance degradation.

TABLE III

ABLATION STUDY OF **DEPTHNET_M** WITH DIFFERENT DECOUPLING NUMBERS K FOR DEPTH ESTIMATION.

Model	AbsRel ↓	SqRel ↓	RMSE ↓	RMSE log ↓
K=2	0.0200	0.0029	0.0946	0.0332
K=3	0.0208	0.0028	0.0955	0.0335
K=4	**0.0201**	**0.0026**	**0.0911**	**0.0326**
K=5	0.0204	0.0028	0.0956	0.0331
K=6	0.0211	0.0031	0.1008	0.0349

TABLE IV

ABLATION STUDY OF **DEPTHNET_M** WITH DIFFERENT CONSTRAINTS FOR DEPTH ESTIMATION. $K = 4$.

Model	AbsRel ↓	SqRel ↓	RMSE ↓	RMSE log ↓
Enable All	**0.0201**	**0.0026**	**0.0911**	**0.0326**
No L_{m}	0.0243	0.0042	0.1173	0.0365
No L_{hier}	0.0197	0.0027	0.0936	0.0327
No L_{sd}	0.0217	0.0031	0.1019	0.0339

Manifold Constraint. While Table I compares the performance of DepthNet and DepthNet_M, Table IV details the contribution of each constraint. The complete removal of both the Hierarchical and Semantic Diversity constraints prevents effective feature decoupling, resulting in a significant performance drop. When only the hierarchical constraint is removed, the model retains its capacity to generate diverse sub-features. However, without hierarchical alignment between global and local semantics, semantic inconsistency arises, leading to information loss and biased feature representations, ultimately degrading overall performance.

IV. CONCLUSIONS

Inspired by advances in manifold learning, this work models endoscopic images as compositions of multiple semantic subspaces embedded within a low-dimensional manifold. To address complex feature representation, we introduce a semantic decoupling perspective. Our core contribution is the Manifold Geometry-Based Feature Decoupling Module (MANDE), which decouples the backbone network's feature map into multiple semantically distinct sub-feature maps guided by novel manifold constraints. Building on MANDE, we propose the Semantic Feature Decoupling-and-Processing Framework (SFDP), a divide-and-conquer architecture for pixel-level analysis. Within SFDP, the decoupled sub-features are processed in parallel by independent task heads. Extensive experiments validate the robustness and generalizability

of our approach. When integrated into SFDP, DepthNet and SwinMTL achieved RMSE reductions of 11.28% and 17.09%, respectively, for depth estimation. Similarly, LSSNet and VM-UNet achieved average MAE reductions of 14.65% and 6.34% for segmentation. These significant improvements were achieved with a parameter increase of only 5.10M, confirming our framework's efficiency and reliability in enhancing generalization for challenging endoluminal scenarios. In summary, our method provides a novel perspective for image understanding in endoscopic vision. By decoupling features into distinct semantic subspaces, it enables task networks to process information in separated domains, thereby effectively reducing feature coupling and substantially boosting perceptual performance in complex endoluminal environments.

V. Acknowledgement

This work was supported by the Science and Technology Commission of Shanghai Municipality (Nos.24511104100), National Key Research and Development Program of China (2025YFC2426300), National Nature Science Foundation of China Grants (82572314, 62477031, 62271246). This manuscript was partially assisted by DeepSeek and ChatGPT for the preparation of the Method section and Experiments section, specifically in translation and language refinement.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[2] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of u-net," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[3] J. Ruan, J. Li, and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation," 2024.

[4] W. Wang, H. Sun, and X. Wang, "Lssnet: A method for colon polyp segmentation based on local feature supplementation and shallow feature supplementation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 446–456.

[5] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," Springer, pp. 205–218, 2022.

[6] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," 2021.

[7] F. Mahmood and N. J. Durr, "Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy," *Medical image analysis*, vol. 48, pp. 230–243, 2018.

[8] M. Meilă and H. Zhang, "Manifold learning: What, how, and why," *Annual Review of Statistics and Its Application*, vol. 11, no. 1, pp. 393–417, 2024.

[9] D. Konstantinidis, I. Papastratis, K. Dimitropoulos, and P. Daras, "Multi-manifold attention for vision transformers," pp. 123 433–123 444, 2023.

[10] Y. Fei, Y. Liu, C. Jia, Z. Li, X. Wei, and M. Chen, "A survey of geometric optimization for deep learning: from euclidean space to riemannian manifold," *ACM Computing Surveys*, vol. 57, no. 5, pp. 1–37, 2025.

[11] P. Mettes, M. Ghadimi Atigh, M. Keller-Ressel, J. Gu, and S. Yeung, "Hyperbolic deep learning in computer vision: A survey," *International Journal of Computer Vision*, vol. 132, no. 9, pp. 3484–3508, 2024.

[12] Y.-W. Luo, C.-X. Ren, D.-Q. Dai, and H. Yan, "Unsupervised domain adaptation via discriminative manifold propagation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1653–1669, 2020.

[13] V. Khrulkov, L. Mirvakhabova, E. Ustinova, I. Oseledets, and V. Lempitsky, "Hyperbolic image embeddings," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6418–6428.

[14] A. Sinha, S. Zeng, M. Yamada, and H. Zhao, "Learning structured representations with hyperbolic embeddings," *Advances in Neural Information Processing Systems*, vol. 37, pp. 91 220–91 259, 2024.

[15] H. Huang, Y. Huang, S. Xie, L. Lin, R. Tong, Y.-W. Chen, Y. Li, and Y. Zheng, "Combinatorial cnn-transformer learning with manifold constraints for semi-supervised medical image segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, 2024, pp. 2330–2338.

[16] M. G. Atigh, J. Schoep, E. Acar, N. Van Noord, and P. Mettes, "Hyperbolic Image Segmentation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 4443–4452.

[17] W. Qian, H. Luo, S. Peng, F. Wang, C. Chen, and H. Li, "Unstructured feature decoupling for vehicle re-identification," in *European Conference on Computer Vision*. Springer, 2022, pp. 336–353.

[18] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *Advances in neural information processing systems*, vol. 30, 2017.

[19] E. Cetin, B. Chamberlain, M. Bronstein, and J. J. Hunt, "Hyperbolic deep reinforcement learning," *arXiv preprint arXiv:2210.01542*, 2022.

[20] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[21] J.-W. Bian, H. Zhan, N. Wang, Z. Li, L. Zhang, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised Scale-Consistent Depth Learning from Video," *International Journal of Computer Vision*, vol. 129, no. 9, pp. 2548–2564, Sep. 2021.

[22] S. Li, W. Lin, Q. Xiang, Y. Tu, S. Asu, and Z. Li, "Unsupervised Photometric-Consistent Depth Estimation from Endoscopic Monocular Video," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 5, pp. 4923–4931, Apr. 2025.

[23] Z. Xie, Z. Geng, J. Hu, Z. Zhang, H. Hu, and Y. Cao, "Revealing the dark secrets of masked image modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14 475–14 485.

[24] P. Taghavi, R. Langari, and G. Pandey, "Swinmtl: A shared architecture for simultaneous depth estimation and semantic segmentation from monocular camera images," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 4957–4964.

[25] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2020, pp. 263–273.

[26] Taylor L Bobrow, Mayank Golhar, Rohan Vijayan, Venkata S Akshintala, Juan R Garcia, and Nicholas J Durr, "Colonoscopy 3D video dataset with paired depth from 2D-3D registration," *Medical Image Analysis*, vol. 90, p. 102956, Dec. 2023.

[27] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *Journal of healthcare engineering*, vol. 2017, no. 1, p. 4037190, 2017.

[28] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *International conference on multimedia modeling*. Springer, 2019, pp. 451–462.

[29] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognition*, vol. 45, no. 9, pp. 3166–3182, 2012.

[30] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *International journal of computer assisted radiology and surgery*, vol. 9, no. 2, pp. 283–293, 2014.

[31] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, vol. 27, 2014.

[32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.