Task Prototype-Based Knowledge Retrieval for Multi-Task Learning from Partially Annotated Data

Youngmin Oh[1,2], **Hyung-Il Kim**[3], **Jung Uk Kim**[1*]
[1]Kyung Hee University
[2]Electronics and Telecommunications Research Institute (ETRI)
[3]Chonnam National University
youngmin@etri.re.kr, hyungil.kim@jnu.ac.kr, ju.kim@khu.ac.kr

Abstract

Multi-task learning (MTL) is critical in real-world applications such as autonomous driving and robotics, enabling simultaneous handling of diverse tasks. However, obtaining fully annotated data for all tasks is impractical due to labeling costs. Existing methods for partially labeled MTL typically rely on predictions from unlabeled tasks, making it difficult to establish reliable task associations and potentially leading to negative transfer and suboptimal performance. To address these issues, we propose a prototype-based knowledge retrieval framework that achieves robust MTL instead of relying on predictions from unlabeled tasks. Our framework consists of two key components: (1) a task prototype embedding task-specific characteristics and quantifying task associations, and (2) a knowledge retrieval transformer that adaptively refines feature representations based on these associations. To achieve this, we introduce an association knowledge generating (AKG) loss to ensure the task prototype consistently captures task-specific characteristics. Extensive experiments demonstrate the effectiveness of our framework, highlighting its potential for robust multi-task learning, even when only a subset of tasks is annotated.

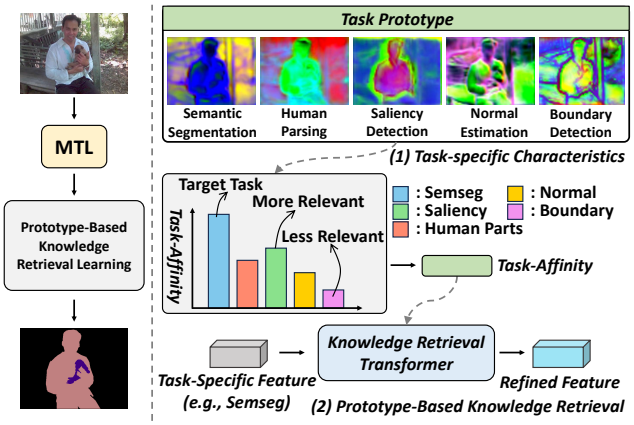

Figure 1: Conceptual diagram of the proposed method (semantic segmentation example). (1) Task prototype generates task-affinity score, and (2) prototype-based knowledge retrieval process utilizes this task-affinity to adaptively enhance task performance.

Introduction

Multi-tasking in computer vision is an important challenge for deploying real-world applications such as autonomous driving (Geiger, Lenz, and Urtasun 2012; Yurtsever et al. 2020) or robotics (Devin et al. 2017), which require a unified process to handle various functional roles (Hu et al. 2023). To this end, multi-task learning (MTL) (Brüggemann et al. 2021; Gao et al. 2019; Lu et al. 2017; Misra et al. 2016; Ye and Xu 2022a; Fan et al. 2022; Liu, Johns, and Davison 2019; Ye and Xu 2023, 2022b) has emerged as a solution, enabling the simultaneous learning of multiple tasks. Unlike traditional single-task learning approaches that train each task independently, the MTL leverages shared associations among tasks (Zamir et al. 2018), facilitating robust predictions across various tasks. By doing so, it has shown remarkable success, particularly in dense prediction tasks (*e.g.,* semantic segmentation and depth estimation).

However, annotating all the tasks across diverse real-world scenarios requires substantial human effort and computational cost during pre-processing for multi-task learning. To mitigate this, recent methods have been proposed to enable robust multi-task learning from partially annotated data. This task is called Multi-Task Partially Supervised Learning (MTPSL). Recent MTPSL works mainly focus on leveraging unlabeled task predictions by utilizing cross-task regularization through joint task-space mappings (Li, Liu, and Bilen 2022) or employing diffusion models combined with multi-task conditioning (Ye and Xu 2024) to integrate cross-task information. However, a common limitation of these approaches is their reliance on *predictions from the unlabeled tasks* when learning target task with available labels. This reliance makes task associations less reliable, as unlabeled tasks often contain noisy or incomplete information, potentially leading to negative knowledge transfer.

To address these limitations, we propose a novel framework for MTPSL that goes beyond simply utilizing pseudo-labels for unlabeled tasks. Our approach introduces a new perspective by explicitly modeling inter-task relationships through task-inherent characteristics, independent of label availability. Ours mainly focuses on capturing inter-task relationships by identifying the inherent characteristics of each

*Corresponding author.

task for more generalized multi-task learning. To this end, we tackle two key challenges: (*i*) how to quantify task associations and (*ii*) how to leverage these associations to adaptively guide reliable knowledge transfer to the target task.

Building upon these two key aspects, as shown in Figure 1, the proposed framework consists of a task prototype and knowledge retrieval transformer. First, for the aspect (*i*), we introduce a task prototype designed to embed task-specific characteristics essential for quantifying task associations. Using this prototype, we generate a task-affinity score to represent the degree of enhancement needed for the target task, based on the associations among tasks. To guide this process, we introduce an association knowledge generating (AKG) loss. It encourages the task prototype to keep task-specific characteristics and learn task-affinity—the degree of enhancement required for each task based on its association with the target task. This task-affinity allows our framework to effectively apply association knowledge, enabling a clear understanding of the enhancement required for transferring to the target task.

Next, to address the aspect (*ii*), we propose a knowledge retrieval transformer that utilizes the task-affinity score as guidance to adaptively perform operations for each task. We generate the task-affinity feature by integrating the task-affinity score with the task prototype, which helps the model retrieve association knowledge needed for enabling enhancements aligned with the target task. Through this feature, each transformer block captures the necessary knowledge to adaptively refine the task-specific feature representations, aligning them closely with the specific requirements of the target task. Based on this, we introduce prototype-based knowledge retrieval learning that enables adaptive enhancement for multi-tasking without relying on predictions from unlabeled tasks. As a result, our approach outperforms the existing state-of-the-art multi-tasking methods, even when only a subset of tasks is annotated.

The major contributions of our paper are as follows:

- We propose a task prototype that captures task-specific features and measures the required enhancement through task associations.
- We develop a knowledge retrieval transformer that uses the task-affinity score to adaptively refine feature representations, aligning them with the specific requirements of the target task.
- We introduce a prototype-based knowledge retrieval learning method that leverages task-specific characteristics instead of relying on predictions from unlabeled tasks. This enhances the performance of diverse tasks, even when annotations are not provided for all tasks.

Related Works

Multi-Task Learning

Multi-task learning have been focused to develop models capable of simultaneously addressing multiple tasks within a single framework. Existing methods (Liu, Johns, and Davison 2019; Gao et al. 2019; Lu et al. 2017; Misra et al. 2016) focus on designing encoder architectures that enable interaction among multiple tasks. Recently, methods (Brüggemann et al. 2021; Ye and Xu 2022a; Yang et al. 2024; Ye and Xu 2023, 2022b; Xu et al. 2018; Fan et al. 2022; Lin et al. 2025) focused on effectively handle multiple tasks using shared features from pre-trained backbone network. For example, ATRC (Brüggemann et al. 2021) explores task relationships to optimize contextual information for multi-task learning. InvPT (Ye and Xu 2022a) introduces a inverted pyramid multi-task transformer that leverages multi-scale feature aggregation for high-resolution task-specific predictions. In (Yang et al. 2024), MLoRE is introduced to explicitly model global task relationships for multi-task dense prediction, and MTMamba (Lin et al. 2025) captures long-range spatial relationships achieves cross-task correlations for multi-task learning. Despite their effectiveness, these methods rely on the assumption that all training samples are fully annotated for every task, which limits their applicability in scenarios where annotations for certain tasks are sparse or unavailable.

Partially Annotated Multi-Task Learning

Recently, several approaches have been proposed to address multi-task learning with partially annotated data, aiming to effectively train models despite incomplete annotations. Annotating all tasks across diverse real-world scenarios incurs significant human and computational costs, making impractical for many applications. To address this, Li *et al.* (Li, Liu, and Bilen 2022) proposed a multi-task partially supervised learning (MTPSL) framework with partially annotated data, introducing cross-task regularization through joint task-space mapping defined for each task pair. DiffusionMTL (Ye and Xu 2024) introduces a diffusion model with multi-task conditioning to improve noisy predictions.

Despite these advances, existing approaches commonly rely on predictions from unlabeled tasks to account for task association. However, the absence of labels can lead to inaccurate predictions, resulting in challenges when utilizing task association effectively. In contrast, our proposed method embeds task-specific characteristics and captures task association instead of relying on predictions from unlabeled tasks. Therefore, the proposed method can be more robust and reliable framework for multi-task learning.

Proposed Method

Figure 2 shows the overall framework of the proposed method, which consists of two parts: multi-task learning and prototype-based knowledge retrieval learning, trained in an end-to-end manner. In multi-task learning, a backbone network receives an input image I to generate an encoded feature f^e, which refined through a vector quantization. This feature passes through the task-specific decoder to generate task-specific feature f^t, where t denotes each particular task (*e.g.,* semantic segmentation or depth estimation). In prototype-based knowledge retrieval learning, f^t passes through the task prototype $\mathcal{V}$ and knowledge retrieval transformer. The task prototype $\mathcal{V}$ generates task-affinity score $\mathcal{A}(\hat{f}^t, \mathcal{V})$, which indicates the degree of enhancement needed for the target task. In the subsequent process, the knowledge retrieval transformer utilizes the task-affinity score as guidance to integrate with the task proto-

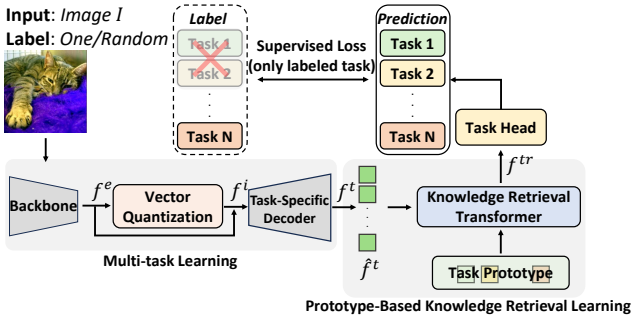

Figure 2: Overview of the proposed multi-task learning with prototype-based knowledge retrieval. It consists of two main components: (1) multi-task learning module, and (2) prototype-based knowledge retrieval module using task prototype and knowledge retrieval transformer.

type $\mathcal{V}$, generating the task-affinity feature f^{ta}. This serves as a key element in guiding the retrieval of task association knowledge and adaptively regulates the enhancement, resulting in the task-refined feature f^{tr}. Finally, each task head uses this f^{tr} to generate task predictions for its corresponding task. More details are in the following subsections.

Vector Quantization for Enhanced Representation

In a partially annotated multi-task setting each task observes only a subset of labels. Unlabeled tasks must still benefit from the others, which requires a shared feature space that is wide enough to hold diverse task cues. We enlarge this space by mapping encoded features to entries in a learnable codebook through vector quantization.

The codebook $\mathcal{Z}$ consists of K learnable slots, defined as $\mathcal{Z} = \{z_k\}_{k=1}^{K}$, with each embedding $z_k \in \mathbb{R}^{1\times c}$ (c represents the dimension of each slot). The encoded feature $f^e \in \mathbb{R}^{h\times w\times c}$ pass through the codebook $\mathcal{Z}$ to generate quantized feature $f^q \in \mathbb{R}^{h\times w\times c}$, by conducting element-wise quantization process $\mathbf{q}(\cdot)$, calculated as:

$$f^q = \mathbf{q}(f^e) := \left(\arg\min_{z_k \in \mathcal{Z}} \|f^e_{ij} - z_k\| \right), \tag{1}$$

where f^e_{ij} denotes the element of the encoded feature.

Next, the quantized feature f^q is integrated with the encoded feature f^e through element-wise summation to obtain the integrated feature f^i. To effectively enhance the shared representation across diverse tasks, we introduce the task-agnostic enhancement loss (TAE) loss, where f^i is passed through a convolutional decoder to reconstruct the input image I^r. $\mathcal{L}_{\text{tae}}$ is formulated as follows:

$$\mathcal{L}_{tae} = \begin{cases} 0.5|I^r - I|^2, & \text{if } |I^r - I| < 1, \\ |I^r - I| - 0.5, & \text{otherwise.} \end{cases} \tag{2}$$

Through $\mathcal{L}_{tae}$, the codebook $\mathcal{Z}$ enhances the shared representation by reconstructing the input image, allowing it to effectively capture task-specific characteristics even when task labels are only partially available.

Task Prototype

In multi-task learning, each task has task-specific characteristics that are essential for leveraging task associations and guiding adaptive enhancements. For instance, human parsing focuses on specific body parts, whereas tasks such as depth estimation or normal estimation require attention to the entire scene. However, predictions from unlabeled tasks generally lack these characteristics compared to those from labeled predictions, making it difficult to understand task-specific characteristics to effectively utilize task associations. Therefore, we propose task prototype $\mathcal{V}$ that quantifies task associations and embeds task-specific characteristics.

Figure 3(a) shows the training process of embedding task knowledge into the task prototype $\mathcal{V}$. The task prototype $\mathcal{V}$ consists of T learnable slots, denoted as $\mathcal{V} = \{v_\tau\}_{\tau=1}^{T} (v_\tau \in \mathbb{R}^{1\times d})$, where T and d indicate the total number of tasks and the dimensionality of each slot, respectively. To capture the task-specific characteristics of each task τ, we first generate task-similarity $S(\hat{f}^t, \mathcal{V}) \in \mathbb{R}^{hw\times T}$, representing the association between target task and the embedded task knowledge of the prototype. Using the task-specific feature $f^t \in \mathbb{R}^{h\times w\times c}$, we apply a convolution layer and flattening to generate $\hat{f}^t \in \mathbb{R}^{hw\times d}$, which is then used with task prototype $\mathcal{V}$ to obtain task-similarity $S(\hat{f}^t, \mathcal{V})$, calculated as:

$$S(\hat{f}^t, \mathcal{V}) = \left(\frac{\hat{f}^t \cdot v_\tau}{\|\hat{f}^t\| \|v_\tau\|} \right)_{\tau=1}^{T}. \tag{3}$$

To embed task knowledge required for reliable transfer to the target task, we introduce the task knowledge embedding (TKE) loss using task-similarity $S(\hat{f}^t, \mathcal{V})$. To achieve this, we perform the softmax function on similarity, generating task-affinity score $\mathcal{A}(\hat{f}^t, \mathcal{V})$. Ideally, the affinity score should be highest for the slot of $\mathcal{V}$ corresponding to the target task t, $\mathcal{L}_{\text{tke}}$ is defined as:

$$\mathcal{L}_{\text{tke}} = -\sum_{t=1}^{T} Y_t \log(\mathcal{A}(\hat{f}^t, \mathcal{V})), \tag{4}$$

where Y_t is a one-hot vector corresponding to target task t. Through $\mathcal{L}_{\text{tke}}$, each task prototype v_τ can memorize task-specific characteristics and capture task association needed for enhancement.

As we explicitly store task-specific characteristics into the task prototype $\mathcal{V}$ using labeled predictions, each v_τ represents the task-specific characteristics of each task. Within the task prototype, task-specific characteristics must remain clearly distinct from those of other tasks while being consistently maintained across all scenarios. To this end, we introduce task consistency (TC) loss using task-specific features from all samples within a batch (see Figure 4). At this time, since our method can generate task-specific features regardless of labels, we newly denote $\hat{f}^t$ as $\hat{x}^t \in \mathbb{R}^{B\times hw\times d}$. Subsequently, we aggregate task-specific features across all samples corresponding to the target task, generating $\tilde{x}^t \in \mathbb{R}^{hw\times d}$. $\mathcal{L}_{\text{tc}}$ focuses on the relationships between task-specific characteristics, ensuring that the accurate characteristics obtained from labeled data remain con-

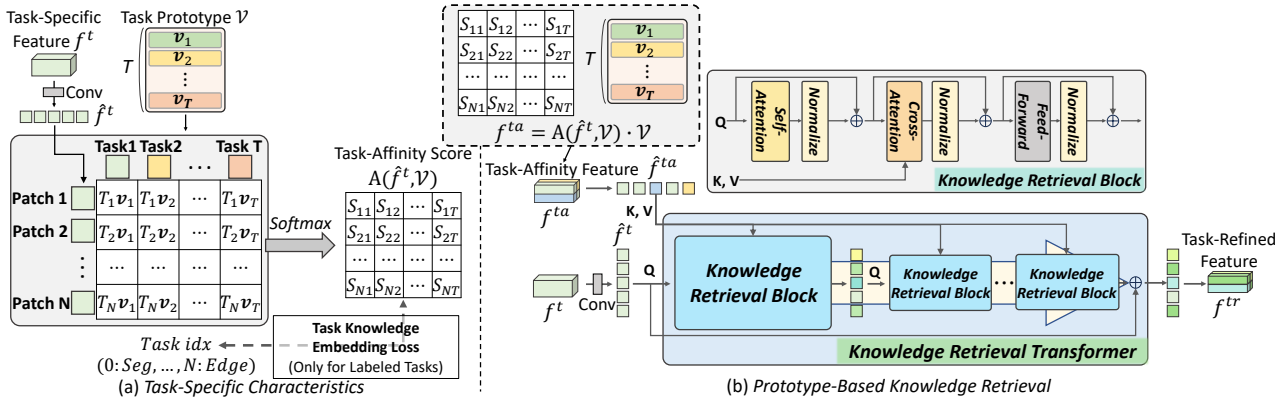

Figure 3: (a) Illustration of the training process for embedding task-specific characteristics into the task prototype. (b) Illustration of the prototype-based knowledge retrieval process.

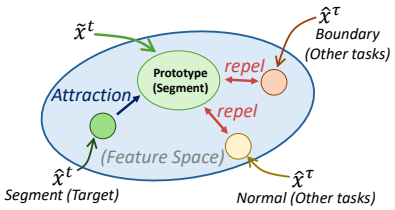

Figure 4: Illustration of the Task Consistency (TC) loss.

sistent across all scenarios of the target task t while remaining clearly separated from those of other tasks τ, defined as:

$$\mathcal{L}_{\text{tc}} = \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{\substack{\tau=1\\ \tau\neq t}}^{T} \max\left(S(\tilde{x}^t, \hat{x}_i^t) - S(\tilde{x}^t, \hat{x}_i^\tau) + \alpha, 0\right), \tag{5}$$

where $S(\cdot,\cdot)$ denotes the cosine similarity.

Finally, the association knowledge generating (AKG) loss $\mathcal{L}_{\text{akg}}$ is obtained by adding $\mathcal{L}_{\text{tke}}$ and $\mathcal{L}_{\text{tc}}$, calculated as:

$$\mathcal{L}_{akg} = \mathcal{L}_{\text{tke}} + \mathcal{L}_{\text{tc}}. \tag{6}$$

In the training phase, the weight parameters of embedding T slots of task prototype $\mathcal{V}$ are initialized randomly and updated through Eq. 6. In the inference phase, all parameters of $\mathcal{V}$ are fixed to recall the consistent task knowledge across all scenarios, generating task-affinity scores that are adaptive to each task without utilizing predictions from unlabeled tasks.

Prototype-Based Knowledge Retrieval

Through task prototype in Section , we generate the task-affinity score $\mathcal{A}(\hat{f}^t, \mathcal{V})$, indicating the association of target task with task-specific characteristics. The core of this section is leveraging these association to adaptively regulate the enhancement to perform transition operations suited for each task. To achieve this, we propose a prototype-based knowledge retrieval method that applies task knowledge from the prototype $\mathcal{V}$, aligning with the requirements of each task.

Figure 3(b) shows the proposed prototype-based knowledge retrieval process, composed of a knowledge-retrieval transformer. This transformer consists of multiple knowledge-retrieval blocks, containing self-attention, cross-attention, and feed-forward network (FFN) layers. To effectively retrieve task knowledge, we first generate the task-affinity feature f^{ta}, which plays a key role in adaptively regulating the enhancement required for each task. As shown in Figure 3(b) (dotted box), f^{ta} is obtained through the matrix multiplication of the task-specific score $\mathcal{A}(\hat{f}^t, \mathcal{V})$ and the task prototype $\mathcal{V}$, which can be represented as:

$$f^{ta} = \mathcal{A}(\hat{f}^t, \mathcal{V}) \cdot \mathcal{V}. \tag{7}$$

After obtaining the task-affinity feature f^{ta}, the knowledge-retrieval block receives the flattened task-specific feature $\hat{f}^t$ and the task-affinity feature $\hat{f}^{ta}$, generating f^t_{ca}. The enhancement is adaptively regulated through the cross-attention layer in each knowledge-retrieval block, where f^{ta} is used as key and value, while f^t (after passing through self-attention layer) serves as query, formulated as:

$$f^t_{\text{sa}} = \text{SelfAtt}(\hat{f}^t), \tag{8}$$

$$f^t_{\text{ca}} = \text{CrossAtt}(f^t_{\text{sa}}, \hat{f}^{ta}, \bar{f}^{ta}). \tag{9}$$

Finally, f^t_{ca} is passed through a feed-forward network, generating the task-refined feature f^{tr}. With the task-affinity feature f^{ta}, our model utilizes task-specific characteristics embedded in the prototype to retrieve task association knowledge as task-affinity, determining the degree of enhancement required for the target task. Subsequently, the cross-attention layer adaptively improves task-specific feature representations by utilizing the association knowledge. This allows our prototype-based knowledge retrieval learning effectively handles diverse tasks without utilizing predictions from unlabeled tasks.

Total Loss Function

The total loss function of our framework is represented as:

$$\mathcal{L}_{Total} = \mathcal{L}_{MTL} + \lambda_1 \sum_{i=1}^{N} \mathcal{L}_{tae} + \lambda_2 \mathcal{L}_{akg}, \tag{10}$$

Method	One Label					Random Label				
	Semseg mIoU↑	**Parsing** mIoU↑	**Saliency** maxF↑	**Normal** mErr↓	**Boundary** odsF↑	**Semseg** mIoU↑	**Parsing** mIoU↑	**Saliency** maxF↑	**Normal** mErr↓	**Boundary** odsF↑
Single-Task Learning	50.34	59.05	77.43	16.59	64.40	51.51	57.90	80.30	15.24	67.80
MTL Baseline	44.73	57.03	75.69	16.47	64.30	46.49	55.39	78.39	15.36	66.80
Semi-Supervised Learning (CVPR'22)	45.00	54.00	61.70	16.90	62.40	59.00	55.80	64.00	15.90	66.90
MTPSL* (CVPR'22)	55.08	56.72	77.06	16.93	63.70	62.44	55.81	78.56	15.45	66.80
DiffusionMTL (Prediction) (CVPR'24)	<u>59.43</u>	56.79	77.57	16.20	64.00	<u>63.68</u>	55.84	79.87	15.38	66.80
DiffusionMTL (Feature) (CVPR'24)	57.78	<u>58.98</u>	<u>77.82</u>	<u>16.11</u>	<u>64.50</u>	62.55	<u>56.84</u>	<u>80.44</u>	<u>14.85</u>	<u>67.10</u>
Proposed Method	**59.78**	**59.08**	**78.62**	**15.63**	**65.10**	**64.30**	**56.87**	**80.51**	**14.48**	**67.30**

Table 1: Quantitative comparison of state-of-the-art MTPSL methods on PASCAL-Context dataset. The results include methods for partially annotated data, along with one and random label settings. * indicates performance reproduced using the same backbone as in (Ye and Xu 2024). **Bold**/<u>underlined</u> fonts indicate the best/second-best results.

where $\mathcal{L}_{MTL}$ denotes supervised loss for multi-task learning with labeled data. It employs the cross-entropy loss for semantic segmentation, human parsing, saliency, and boundary detection, while the L1-norm loss is used for depth and surface normal estimation. λ denotes balancing parameter.

Experiments

Dataset and Evaluation Metrics

PASCAL-Context. PASCAL-Context (Everingham et al. 2010) dataset contains 4,998 training images and 5,105 testing images, which provide annotations for dense prediction tasks such as semantic segmentation, human parsing, and object boundary detection. Additionally, pseudo labels (Maninis, Radosavovic, and Kokkinos 2019) for tasks like surface normal estimation and saliency detection have been generated, making it a comprehensive benchmark for multi-task learning. Following (Li, Liu, and Bilen 2022; Ye and Xu 2024), we utilize all the tasks for the evaluation.

NYUD-v2. NYUD-v2 (Silberman et al. 2012) dataset contains 795 training images and 654 testing images, collected from various indoor scenarios. It includes annotations for 13-class semantic segmentation, depth estimation, and surface normal estimation. Following the protocol of existing works (Li, Liu, and Bilen 2022; Ye and Xu 2024), the resolutions of all images were resized to 288×384.

Evaluation Metrics. To compare the performance under partially annotated settings, we adopted the same protocol as prior research (Li, Liu, and Bilen 2022), where label configurations are predefined. Specifically, two label configurations are used: (*i*) one-label setting, where each training image is annotated for only one task, and (*ii*) random-label setting, where each image is provided with annotations for at least one task and at most a predefined number of tasks.

For evaluation, we use metrics from prior works (Li, Liu, and Bilen 2022; Ye and Xu 2024; Yang et al. 2024). Mean Intersection over Union (mIoU) is used for semantic segmentation and human parsing, while the maximal F-measure (maxF) evaluates saliency detection. For surface normal estimation, we use mean angular error (mErr), and for boundary detection, the optimal-dataset-scale F-measure (odsF). Absolute error (absErr) is employed for depth estimation.

Implementation Details

Following the methods in (Ye and Xu 2024), we use ResNet-18 as our backbone (He et al. 2016). All experiments were conducted on a single RTX A6000 GPU. For both PASCAL-Context and NYUD-v2 datasets, we train our method using the Adam optimizer with an initial learning rate of 2×10^{-5}. We trained the model for 100 epochs with a batch size of 6 on PASCAL-Context, and 200 epochs with a batch size of 4 on NYUD-v2, following previous work (Li, Liu, and Bilen 2022; Ye and Xu 2024). For the codebook and task prototype, we used $K = 4096$ slots for the codebook, and for the task prototype T, we used $T = 5$ for PASCAL-Context and $T = 3$ for NYUD-v2. Each slot has a dimension of 1024, with 8 heads for cross-attention in the knowledge retrieval transformer, and the output feature has 1024 channels. The task-specific decoder consists of 3×3 convolution layers with ReLU, and the task head is a 1×1 convolution layer.

Comparision with the State-of-the-art Methods

Results on the PASCAL-Context Dataset. We compared our method with the state-of-the-art method (Li, Liu, and Bilen 2022; Ye and Xu 2024) under partially annotated settings on the PASCAL-Context dataset. As shown in Table 1, while DiffusionMTL (Ye and Xu 2024) has shown improved performance across all tasks, its performance varied according to input type of the diffusion, *i.e.*, prediction map (prediction) or feature map (feature). In contrast, our method outperforms across all tasks, maintaining consistent performance regardless of the learning strategy.

Results on NYUD-v2. We also compared on NYUD-v2 dataset to demonstrate the generalizability of our method. As shown in Table 2, ours consistently outperforms existing methods. Since our task prototype embeds task-specific characteristics and knowledge retrieval transformer leverages them to adaptively enhance feature representations by capturing task associations, ours shows robust performance.

Ablation Studies

Effect of the Proposed Loss Functions. Table 3 shows the effectiveness of the proposed loss functions. When vector quantization is added with $\mathcal{L}_{tae}$, it is slightly improved by enhancing shared feature representations. This allows the task prototype to better capture task-specific characteristics.

Method	One Label			Random Label		
	Semseg mIoU↑	Depth absErr↓	Normal mErr↓	Semseg mIoU↑	Depth absErr↓	Normal mErr↓
Single-Task Learning	45.28	0.4802	25.93	48.25	0.4792	24.65
MTL Baseline	42.77	0.5134	26.99	44.82	0.4886	25.92
Semi-Supervised Learning (CVPR'22)	27.52	0.6499	33.58	29.50	0.6224	33.31
MTPSL* (CVPR'22)	43.97	0.5140	26.30	46.03	0.4811	25.97
DiffusionMTL (Prediction) (CVPR'24)	<u>44.97</u>	0.5137	26.17	<u>47.44</u>	0.4803	25.26
DiffusionMTL (Feature) (CVPR'24)	44.47	<u>0.5059</u>	<u>25.84</u>	46.82	<u>0.4743</u>	<u>24.75</u>
Proposed Method	**45.95**	**0.4865**	**25.64**	**47.53**	**0.4621**	**24.67**

Table 2: Quantitative comparison of state-of-the-art MTPSL methods on NYUD-v2 dataset. * indicates results reproduced using the same backbone. **Bold**/<u>underlined</u> fonts indicate the best/second-best results.

$\mathcal{L}_{tae}$	$\mathcal{L}_{akg}$		Semseg	Parsing	Saliency	Normal	Boundary
	$\mathcal{L}_{tke}$	$\mathcal{L}_{tc}$	mIoU ↑	mIoU ↑	maxF ↑	mErr ↓	odsF ↑
-	-	-	44.73	57.03	75.69	16.47	64.38
✓	-	-	44.83	57.13	76.13	16.22	64.50
✓	✓	-	58.21	58.87	78.50	15.67	65.00
✓	✓	✓	**59.78**	**59.08**	**78.62**	**15.63**	**65.10**

Table 3: Ablation study to investigate the effect of proposed loss functions on PASCAL-Context (one-label setting).

# Dimension	Param	Semseg mIoU ↑	Depth absErr ↓	Normal mErr ↓
-	146.5M	42.77	0.5134	26.99
256	156.6M	44.91	0.4931	25.67
512	157.5M	45.65	0.4878	25.73
1024	159.4M	**45.95**	**0.4865**	**25.64**
2048	163.0M	45.33	0.4867	25.77

Table 4: Effect of task prototype on NYUD-v2 under the one label setting by varying the dimension of its slot T.

In $\mathcal{L}_{akg}$, we embed task-specific characteristics into the task prototype using $\mathcal{L}_{tke}$, enabling the prototype-based knowledge retrieval module to effectively capture task associations. This enhances feature representations, outperforming the baseline. With $\mathcal{L}_{tc}$, the task prototype effectively ensures consistency in task-specific characteristics and demonstrates superior performance across all tasks.

Effect of dimensionality of task prototype slot T. We conducted experiments by varying the dimensionality of task prototype slots T. As shown in Table 4, when the dimensionality is small, fewer parameters make it difficult to embed sufficient task-specific characteristics, reducing performance. Conversely, a large dimensionality makes it struggle to utilize the information, also leading to degraded performance. Optimal performance was achieved at 1024.

Visualization Results

Figure 5 shows the qualitative comparisons between our method and DiffussionMTL (Ye and Xu 2024) on PASCAL-Context under the one-label setting. Existing method shows lower quality on certain tasks (*e.g.,* semantic segmentation), while our method achieves improved results across all tasks.

Method	Prompting Method	NYUD-v2		
		Semseg mIoU ↑	Depth absErr ↓	Normal mErr ↓
MTL Baseline ($\mathcal{B}$)	-	45.41	0.4277	22.34
$\mathcal{B}$+TaskPrompter (ICLR'22)	Latent Learn.	48.68	0.4141	20.65
$\mathcal{B}$+TSP-Transformer (WACV'24)	Latent Learn.	46.26	0.4239	21.00
Proposed Method	Explicit Learn.	**50.08**	**0.3857**	**20.57**

Table 5: Comparison with different prompt-based methods on the NYUD-v2 dataset under one label setting. $\mathcal{B}$ is baseline MTL network with a ViT-L backbone. The results are obtained via our reproduction with the official source code.

Discussions

Effect of Prototype-Based Knowledge Retrieval. We investigate the effectiveness of our prototype-based knowledge retrieval learning compared to existing prompt-based multi-task learning methods (Ye and Xu 2022b; Wang et al. 2024). The existing approaches focus on embedding task prompt within the transformer architecture, where prompts are learned in *a supervised manner under fully labeled conditions*. Therefore, as shown in Table 5, while they show improved performance over the baseline, they only work when labeled data is available. In contrast, since we leverage task prototype with $\mathcal{L}_{akg}$, ours outperforms across all tasks, even when only a subset of tasks is annotated.

Visualization of Task Prototype Capabilities. To demonstrate how our method utilizes task associations through the task prototype, we visualize an attention map in Figure 6. The activated slots highlight the affinity for task-specific characteristics required to enhance the target task. By leveraging these associations, our method enables effective knowledge retrieval and enhances feature representations without relying on predictions from unlabeled tasks.

Visualization of Embedding Parameters in Task Prototype. To evaluate how effectively our method captures task-specific characteristics, we visualize the embedding parameters of task prototype in Figure 7. The task prototype contains individual slots for each task, where the elements in each slot represent the task-specific characteristic information. While these characteristics are distinct for each task, certain elements share similar properties across tasks. This

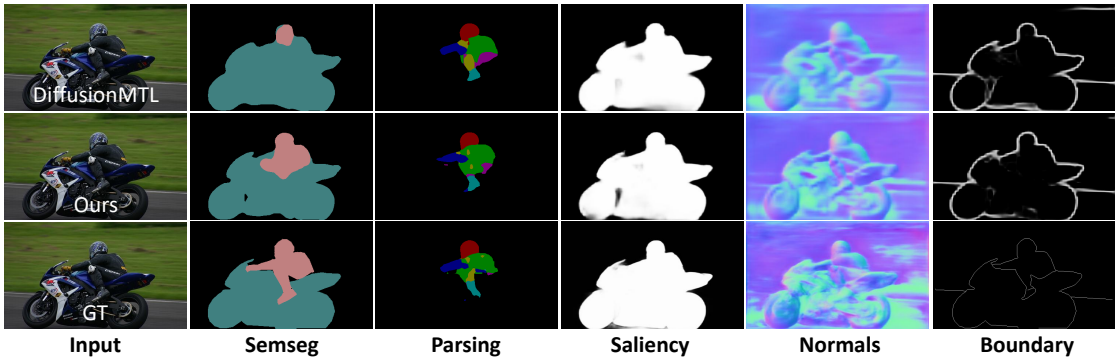

Figure 5: Qualitative comparison of our method and state-of-the-art method on PASCAL-Context under one-label setting.

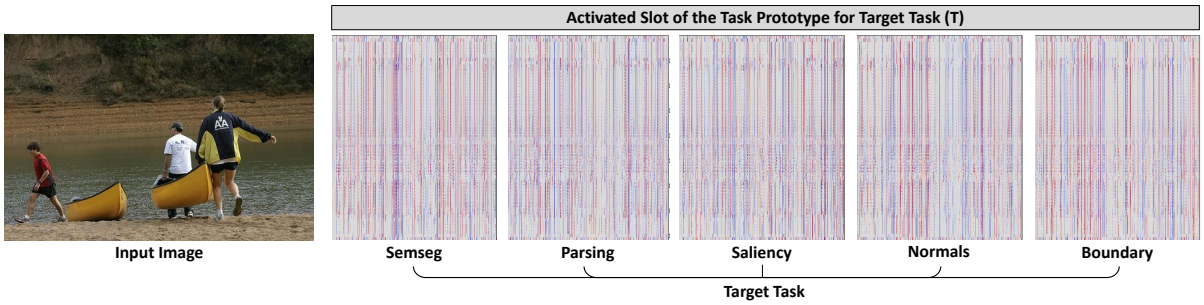

Figure 6: Example of the task prototype. When target tasks are different, the activated slot of the task prototype is different.

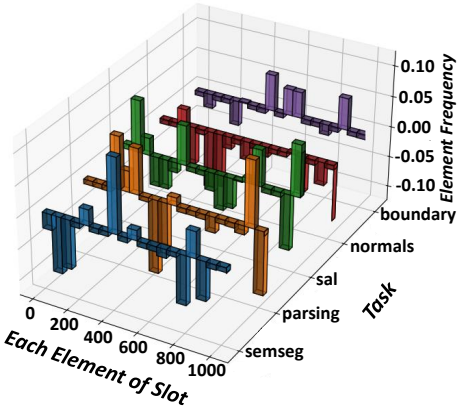

Figure 7: Visualization of task prototype.

allows our method to capture task associations, facilitating effective knowledge retrieval and adaptive enhancements.

Generalization Ability Across Different Backbone. Table 6 shows the generalizability of our method using a different backbone network, ResNet-50. Our method outperforms the others across all tasks, demonstrating its effectiveness and generalizability regardless of the backbone architecture.

Limitations. Although our method captures task-specific characteristics and regulates task associations to enhance multi-task learning without relying on predictions from unlabeled tasks, it is currently designed for tasks seen during training. Extending this framework to unseen tasks in a zero-shot or meta-learning remains an open challenge.

Method	PASCAL-Context				
	Semseg	Parsing	Saliency	Normal	Boundary
	mIoU ↑	mIoU ↑	maxF ↑	mErr ↓	odsF ↑
DiffusionMTL (Prediction)	60.92	59.94	77.58	17.31	63.80
DiffusionMTL (Feature)	58.78	61.91	77.07	16.49	66.20
Proposed Method	**62.23**	**62.14**	**78.10**	**16.19**	**66.70**

Table 6: Comparision of state-of-the-art multi-task learning method with different backbone network on PASCAL-Context dataset under one-label setting.

Conclusion

We introduce a novel framework for prototype-based knowledge retrieval learning, designed to effectively leverage task-specific characteristics and associations without relying on predictions from unlabeled tasks. Addressing the challenges of partially annotated data, we introduce a task prototype with association knowledge generating loss to embed task-specific characteristics and to generate task-affinity score. Also, we propose the knowledge retrieval transformer adaptively enhance feature representations for each task with task prototype. As a result, our method can reliable transfer across all tasks, even without additional annotations.

Acknowledgments

This work was supported by the NRF grant funded by the Korea government (MSIT) (No. RS-2023-00252391), and by IITP grant funded by the Korea government (MSIT) (No. RS-2022-00155911: Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University), No. RS-2023-00236245, Development of Perception/Planning AI SW for Seamless Autonomous Driving in Adverse Weather/Unstructured Environment (25%), No. RS-2022-II220124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities (20%), No. RS-2024-00509257: Global AI Frontier Lab).

References

Brüggemann, D.; Kanakis, M.; Obukhov, A.; Georgoulis, S.; and Van Gool, L. 2021. Exploring relational context for multi-task dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15869–15878.

Devin, C.; Gupta, A.; Darrell, T.; Abbeel, P.; and Levine, S. 2017. Learning modular neural network policies for multi-task and multi-robot transfer. In *2017 IEEE international conference on robotics and automation (ICRA)*, 2169–2176. IEEE.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Fan, Z.; Sarkar, R.; Jiang, Z.; Chen, T.; Zou, K.; Cheng, Y.; Hao, C.; Wang, Z.; et al. 2022. M^{3}vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. *Advances in Neural Information Processing Systems*, 35: 28441–28457.

Gao, Y.; Ma, J.; Zhao, M.; Liu, W.; and Yuille, A. L. 2019. Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3205–3214.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17853–17862.

Li, W.-H.; Liu, X.; and Bilen, H. 2022. Learning multiple dense prediction tasks from partially annotated data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18879–18889.

Lin, B.; Jiang, W.; Chen, P.; Zhang, Y.; Liu, S.; and Chen, Y.-C. 2025. MTMamba: Enhancing multi-task dense scene understanding by mamba-based decoders. In *European Conference on Computer Vision*, 314–330. Springer.

Liu, S.; Johns, E.; and Davison, A. J. 2019. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1871–1880.

Lu, Y.; Kumar, A.; Zhai, S.; Cheng, Y.; Javidi, T.; and Feris, R. 2017. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5334–5343.

Maninis, K.-K.; Radosavovic, I.; and Kokkinos, I. 2019. Attentive single-tasking of multiple tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1851–1860.

Misra, I.; Shrivastava, A.; Gupta, A.; and Hebert, M. 2016. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3994–4003.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, 746–760. Springer.

Wang, S.; Li, J.; Zhao, Z.; Lian, D.; Huang, B.; Wang, X.; Li, Z.; and Gao, S. 2024. Tsp-transformer: Task-specific prompts boosted transformer for holistic scene understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 925–934.

Xu, D.; Ouyang, W.; Wang, X.; and Sebe, N. 2018. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 675–684.

Yang, Y.; Jiang, P.-T.; Hou, Q.; Zhang, H.; Chen, J.; and Li, B. 2024. Multi-Task Dense Prediction via Mixture of Low-Rank Experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27927–27937.

Ye, H.; and Xu, D. 2022a. Inverted pyramid multi-task transformer for dense scene understanding. In *European Conference on Computer Vision*, 514–530. Springer.

Ye, H.; and Xu, D. 2022b. Taskprompter: Spatial-channel multi-task prompting for dense scene understanding. In *The Eleventh International Conference on Learning Representations*.

Ye, H.; and Xu, D. 2023. Taskexpert: Dynamically assembling multi-task representations with memorial mixture-of-experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21828–21837.

Ye, H.; and Xu, D. 2024. DiffusionMTL: Learning Multi-Task Denoising Diffusion Model from Partially Annotated Data. 27960–27969.

Yurtsever, E.; Lambert, J.; Carballo, A.; and Takeda, K. 2020. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8: 58443–58469.

Zamir, A. R.; Sax, A.; Shen, W.; Guibas, L. J.; Malik, J.; and Savarese, S. 2018. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3712–3722.