Going Beyond Multi-Task Dense Prediction with Synergy Embedding Models

Huimin Huang[1], Yawen Huang[2 †], Lanfen Lin[1,*] Ruofeng Tong[1,3], Yen-Wei Chen[4*],
Hao Zheng[2], Yuexiang Li[5], Yefeng Zheng[2]
[1] Zhejiang University, [2] Jarvis Research Center, Tencent YouTu Lab, [3] Zhejiang Lab,
[4] Ritsumeikan University, [5] Guangxi Medical University

Abstract

Multi-task visual scene understanding aims to leverage the relationships among a set of correlated tasks, which are solved simultaneously by embedding them within a unified network. However, most existing methods give rise to two primary concerns from a task-level perspective: (1) the lack of task-independent correspondences for distinct tasks, and (2) the neglect of explicit task-consensual dependencies among various tasks. To address these issues, we propose a novel synergy embedding models (SEM), which goes beyond multi-task dense prediction by leveraging two innovative designs: the intra-task hierarchy-adaptive module and the inter-task EM-interactive module. Specifically, the constructed intra-task module incorporates hierarchy-adaptive keys from multiple stages, enabling the efficient learning of specialized visual patterns with an optimal trade-off. In addition, the developed inter-task module learns interactions from a compact set of mutual bases among various tasks, benefiting from the expectation maximization (EM) algorithm. Extensive empirical evidence from two public benchmarks, NYUD-v2 and PASCAL-Context, demonstrates that SEM consistently outperforms state-of-the-art approaches across a range of metrics.

1. Introduction

Dense scene understanding is a rapidly growing field that learns multiple objectives from shared representations [17, 24, 34], allowing for improving the efficiency and accuracy of each task. Its success on computer vision encompassing numerous dense prediction tasks, for example, semantic segmentation [38, 47], boundary detection [11, 27], and geometric tasks like depth/normal estimation [10, 15, 21]. Intuitively, these dense prediction tasks inherently possess distinctive representations (as illustrated in Fig. 1). This distinction is of significant importance when striving for excellence in resolving individual tasks. On the other hand, exploring the interrelatedness among tasks is also crucial, as leveraging these cross-task synergies can provide mutual benefits. For instance, abrupt alterations in depth maps could indicate semantic boundaries in segmentation maps. Similarly, pixels belonging to certain semantic classes, such as “bed”, might exhibit similar surface normals [22].

In this context, potential advantages of using both task-independent and task-consensual representations go beyond the direct implications of learning independently, since learning multiple related tasks has been empirically shown to often significantly improve performance [42, 43]. However, there remains a challenging issue in effectively learning the unique characteristics of each task (intra-task specialty) and the complementary aspects across separate tasks (inter-task complementarity) within a unified model. As depicted in Fig. 1, the multi-task learning baseline (MTLB) produces subpar predictions, with feature maps for each task tending to be confusing and lacking clear distinction, leading to unsatisfactory results for individual tasks.

In the pursuit of advancing multi-task learning (MTL), most existing research [8, 24, 33, 39] heavily relies on the capabilities of Convolutional Neural Networks (CNN) [20, 32]. Significant progresses have been made in developing multi-task optimization losses [14], as well as in designing [18, 33] or searching [9] for multi-task information sharing strategies and network structures. Despite the success of CNN-based MTL models, which have shown promising performance in multi-task dense prediction tasks, these algorithms are still constrained by the limitations inherent in convolutional operations. Specifically, they lack the ability for global modeling and cross-task interaction [41]. To address these problems, recent Transformer-based MTL methods [1, 40, 41] have leveraged the attention mechanism [6, 19, 36], enabling effective global modeling and task interactions. However, these approaches still struggle with two key issues. **(i) Intra-task dependency**: Existing Transformer-based methods either focus on learning at a single stage or the simply concatenated multi-stage feature maps. Such approaches overlook the importance of each

[*] Corresponding Authors: Lanfen Lin (llf@zju.edu.cn), Yen-Wei Chen (chen@is.ritsumei.ac.jp).

[†] Huimin Huang and Yawen Huang are co-first authors, and this work is done during the internship at Tencent YouTu Lab.

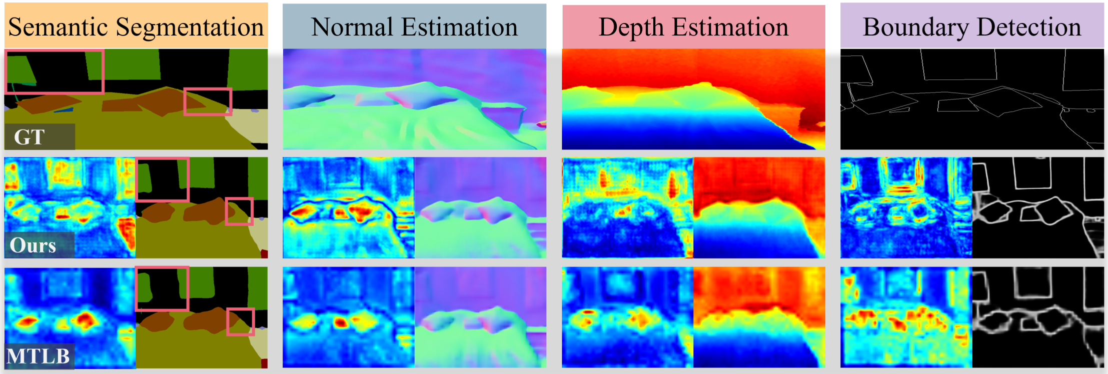

Figure 1. Visualization of the Ground Truths (GT) in the 1^{st} row, the feature maps (left) and predictions (right) of our method (Ours) in the 2^{nd} row and multi-task learning baseline (MTLB) in the 3^{rd} row for four different tasks on the NYUD-v2 dataset. Intuitively, the MTLB suffers from the investigation of task-independent internal structure, leading to the confused feature maps and unsatisfactory results for individual tasks. Impressively, our method efficiently alleviates above limitations and consequently improves the performance from the intra-task and inter-task perspectives, achieving competitive results for multiple scene understanding tasks simultaneously.

stage and the cross-stage correlation during the construction of task-specific features. Moreover, the weighted aggregation across all positions can lead to questionable pixel-wise relationships, potentially undermining the consistent intra-task dependencies. **(ii) Inter-task dependency**: According to a recent survey of Transformer [41], they simply treat all pixels, aggregated from different tasks, as bases for reconstructing the feature space by focusing on pixel-to-pixel dependencies. This may not be an optimal design, as the inherent all-pair dependencies could be muddled and disrupted by inconsistent details, leading to redundant computations and unnecessary noise that hinder high performance.

To overcome these challenges, we propose synergy embedding models (referred to as SEM), a novel architecture powered by both EM-driven learning and hierarchical adaption strategy with task conscious for multi-task dense prediction (summarized in Fig. 2). Firstly, considering the unique characteristics of context modeling at different stages, particularly when establishing task-independent features grounded on multi-task dense predictions, we design an intra-task hierarchy-adaptive module. This module is aimed at exploring task-specific visual patterns, thereby yielding a more consistent representation for each task. Specifically, we leverage features from multiple stages of the encoder and update them individually by applying a Transformer with global reasoning capabilities. However, traditional Transformers involve a large number of keys for each query, leading to redundant computations with high complexity. To address this, we design a lightweight Transformer for the specialized task, equipped with hierarchy-adaptive keys/values. This Transformer adaptively searches for salient points at each stage and measures the importance of each stage, effectively combining detailed aspects (from shallow stages) and semantic aspects (from deep stages) to construct task-specific feature representations.

Secondly, instead of using all pixels as the reconstruction bases, we employ EM algorithms to map a series of approximately compact basis sets from the original over-compact ones, thereby generating a Hilbert space [44]. We use maximum likelihood along with appropriate filtering algorithms to handle noisy observations, leading to accurate parameter estimates with high computational efficiency [16, 25, 29]. Specifically, we construct a task-consensual basis for interaction among tasks. The number of these bases corresponds to the total number of key points, representing a simplified operation compared to the original space.

Our method, SEM, demonstrates its ability to model task consciousness as shown in Fig. 1. Compared to the baseline, SEM enhances the generation of task-specific representations and achieves accurate multi-task predictions by learning mutual information. In summary, our contributions are four-fold: (**1**) We scrutinize the intra-task and inter-task dependencies that existing Transformer-based methods overlook, and propose two plug-and-play modules from a task-level perspective, referred to as SEM, to improve the performance of multi-task dense predictions; (**2**) We introduce an intra-task hierarchy-adaptive module to learn cross-stage dependencies, which aids in generating task-independent representations; (**3**) We design an inter-task EM-based interactive module to learn task-consensual bases, facilitating interaction among different tasks; (**4**) Extensive empirical evidence from two public benchmarks, NYUD-v2 and PASCAL-Context, validates that our method consistently outperforms state-of-the-arts across various metrics.

2. Related Work

Multi-Task Deep Learning. Recently, multi-task deep learning has been validated to enhance the training efficiency of scene understanding tasks. The field primarily revolves around two main paradigms [34, 42]: multi-

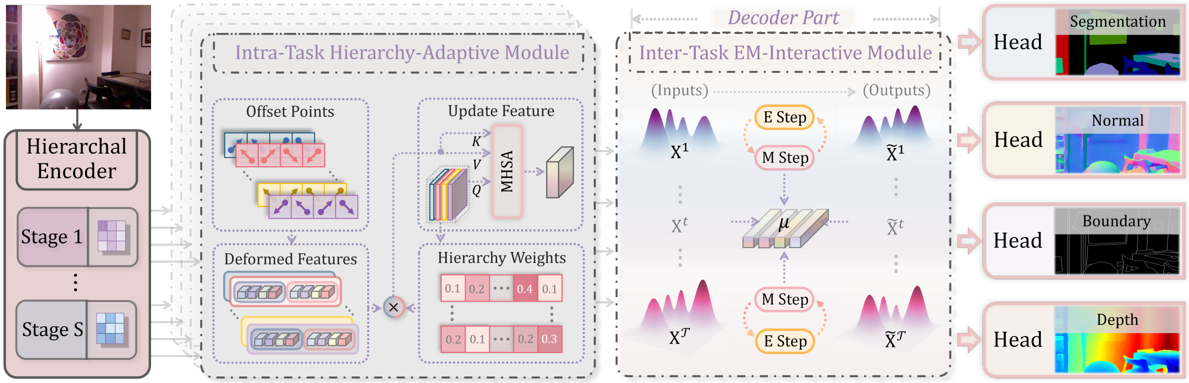

Figure 2. **The overall architecture of SEM.** There are two main task-aware sub-modules: (a) intra-task hierarchy-adaptive module that receives multi-stage representations generated from the encoder and outputs task-independent features for the decoder; (b) inter-task EM-interactive module that is plugged into the decoder stage for learning task-consensual correlations.

task optimization and network structure design. The former paradigm aims to mitigate the issue of task competition by balancing multiple loss optimizations during the training process [4, 5, 14]. The latter paradigm, on the other hand, is extensively explored by investigating architectural designs that facilitate interactions among tasks. Inspired by the superiority of the Transformer [6, 19] in extracting global dependencies, several Transformer-based models have been proposed for multi-task learning [40–43]. For instance, DeMT [41] proposed a multi-task model that leveraged the strengths of both deformable CNN and query-based Transformer. InvPT [42] explored spatial and cross-task relationships at a global level, while TaskPrompter [43] designed a multi-task prompting framework that stimulates spatial-wise and channel-wise prompt learning. Despite their innovative designs, these existing methods still struggle with limitations stemming from insufficient cross-stage relations and redundant cross-task interactions. To address these issues, we propose an effective Transformer-based model that simultaneously captures hierarchical intra-task dependency and EM-driven inter-task dependency.

Expectation Maximization in Computer Vision. With the recent surge in deep learning applications, numerous studies have integrated modern networks with the well-established EM algorithm to harness its clustering and filtering capabilities [7, 13, 31]. For instance, SSN [12] combined EM-based iterations with a neural network to develop efficient superpixel sampling. Inspired by the success of attention scheme, EMANet [16] proposed an EM-based attention method that iteratively generates a more compact basis set, thereby reducing computational complexity. AEMA-Net [45] further extended EMANet into a 3D asymmetric EM-based attention network, which enhanced the performance of brain tumor segmentation. Enlightened by the flexible and accurate parameter estimation in the EM-based algorithm, we design a simple yet effective EM-based interactive module to construct a task-consensual basis among tasks.

3. Method

3.1. Overview of ClassFormer

The proposed SEM is designed to investigate both intra-task hierarchical cues and inter-task correlations, as depicted in Fig. 3. Specifically, we employ the Transformer backbone as the shared encoder for individual tasks to yield stage-wise representations from the cascaded blocks, and thereby exploring long-distance relationships. Leveraging these hierarchical features, the intra-task reasoning (in Sec. 3.2) adaptively explores deformed points across stages. This process decouples the characteristics of various tasks and updates them separately based on their internal task structures. To further understand the correlation among various tasks, we build an inter-task EM-based interactive module (in Sec. 3.3). This module can explore the correspondences among multiple-task features. Finally, the enhanced representations, which incorporate both task-independent and task-consensual contexts, are sent to a specialized head to produce pixel-wise prediction. In the following sections, we will elaborate on the details of our SEM.

3.2. Intra-Task Hierarchy-Adaptive Module

To decouple task-aware independence from the task-agnostic features produced by a shared encoder, our approach identifies the distinctive salient region of each task. This is particularly evident when considering the different stages of representation; for instance, shallow layers tend to capture more detailed, fine-grained features, whereas deep layers increasingly focus on broader semantic context. Given that each task may place varying levels of importance on different stages of representation, it is crucial to establish a mechanism for capturing the long-range dependencies across these stages. A direct approach would be to employ a transformer block on the long sequence formed by concatenating multi-stage tokens. However, this method can lead to unnecessary computational overhead and the potential for

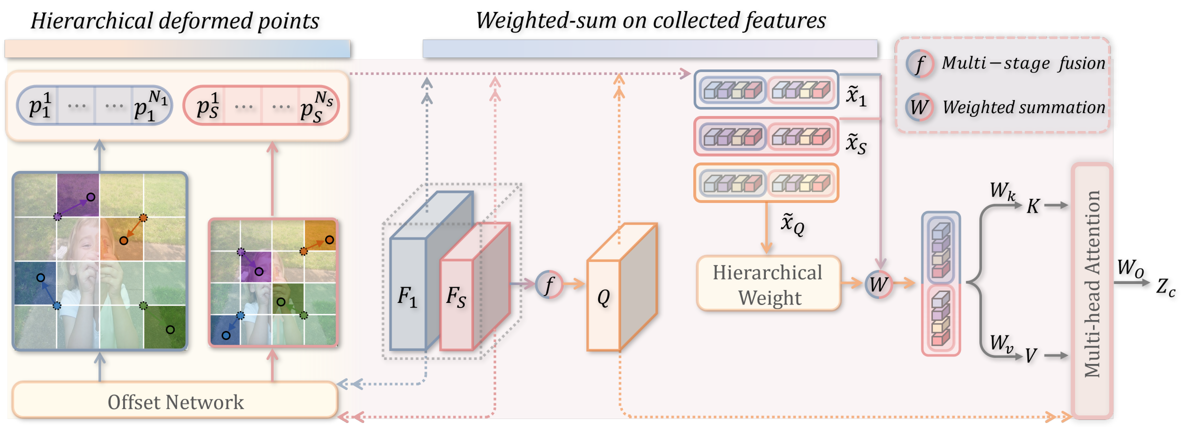

Figure 3. An overview of intra-task hierarchy-adaptive module. For clarity, we show the case with two stages (*i.e.,* $S = 2$) and four deformed points in each stage (*i.e.,* $N_s = 4, s = 1, \cdots, S$).

interference due to the processing of irrelevant correlations. To address this, we design an intra-task transformer that utilizes a hierarchy-adaptive key/value mechanism tailored to the specific task at hand. This allows for the learning of task-relevant visual patterns with a more efficient balance between computational cost and performance.

Hierarchical deformed points. As illustrated in the Fig. 3, given a set of task-agnostic features from S hierarchical stages $F = \{F_1, \dots, F_S\}$, we individually sent F into the intra-task module with hierarchy exploration for $\mathcal{T}$ times (where $\mathcal{T}$ is the number of tasks), which is capable of generating a collection of decoupled feature maps with task-specific cues. To unify the channel-wise dimension of each stage, a convolution operation is first utilized to transfer the channel into the same number of C. Enlightened by [37], a set of uniform grids is considered as reference points, generated from a downsampling rate r_s built upon the input feature $F_s \in \mathbb{R}^{H_s \times W_s \times C}$. In this way, the values of reference points are formed into 2D spatial coordinates $\{(0,0), \cdots, (\lfloor H_s/r_s \rfloor - 1, \lfloor W_s/r_s \rfloor - 1)\}$. Subsequently, a normalization operation is adopted to further constrain the range into $[-1, +1]$.

To generate the corresponding offsets for all reference points, an offset network $\theta_{offset}(\cdot)$ is designed and takes the stage-wise feature F_s as input and outputs $\Delta p = \theta_{offset}(F_s)$. Specifically, the offset network is designed with a convolution with stride r_s, following a GELU activation and another convolution to generate the $\Delta p \in \mathbb{R}^{\lfloor \frac{H_s}{r_s} \rfloor \times \lfloor \frac{W_s}{r_s} \rfloor \times 2}$. Similarly, $\tanh(\cdot)$ is applied on the generated Δp to scale the value into $[-1, +1]$. The locations of deformed points p_s can be obtained by:

$$p_s = p_s^{ref} + \theta_{offset}(x_s). \tag{1}$$

In this way, a collection of deformed points from hierarchical features can be obtained by $p = \{p_s | s = 1, \dots, S\}$. Accordingly, the corresponding features of selected points can be further sampled, by adopting the differentiable bilinear interpolation $\phi(\cdot;\cdot)$:

$$\widetilde{x}_s = \phi(F_s; p_i^j | i = 1, \dots, S; j = 1, \dots, N_s), \tag{2}$$

where N_s is the number of deformed points in the s-th stage, which is equals to $\lfloor H_s/r_s \rfloor \times \lfloor W_s/r_s \rfloor$. Accordingly, the collected features $\widetilde{x} = \{x_s | s = 1, \dots, S\}$ have the size of $S \times N \times C$, and N denotes the total number of hierarchical deformed points, which is equals to $\sum_{s=1}^{S} N_s$.

Hierarchical weights. It is vital to emphasize the importance of each stage, which can optimally highlight the informative features while suppressing redundant feature maps. Enlightened by this, we aim to generate the hierarchical weight $\mathcal{W}$ from the query Q, adaptively. Specifically, embedded with hierarchical cues, Q is generated by unifying the multi-stage features into the same minimum feature dimension (*i.e.,* $H_S \times W_S$), and followed by a 3×3 convolution operation and GELU activation, resulting in $Q \in \mathbb{R}^{H_S \times W_S \times C}$. Then, the deformed features $\widetilde{x}_Q \in \mathbb{R}^{N \times C}$ on the aggregated Q are selected position-wisely, according to the sampling function $\phi(\cdot;\cdot)$ in Eq. (2). Then, a sub-network $\theta_{HW}(\cdot)$ is adopted for measuring the importance of stages for each deformed point. Specifically, $\theta_{HW}(\cdot)$ is designed with two linear operations with a RELU function. The whole process can be formulated as:

$$\mathcal{W} = \theta_{HW}(\widetilde{x}_Q) \in \mathbb{R}^{N \times S}, \quad \text{where } \widetilde{x}_Q = \phi(Q; p). \tag{3}$$

Notably, we apply the softmax function to constrain its value into [0,1]. Benefiting from the adaptive weight $\mathcal{W}$, the hierarchy-aware key k and value v can be obtained by the weighted average of deformed points from multiple stages, followed by two projection matrices W_k and W_v:

$$\widetilde{K} = \left(\mathcal{W}^T \otimes \widetilde{x}\right) W_k, \;\; \widetilde{V} = \left(\mathcal{W}^T \otimes \widetilde{x}\right) W_v. \tag{4}$$

$\otimes$ is the weighted summation operation. By doing this, the aggregated K and V are embedded with the cross-stage

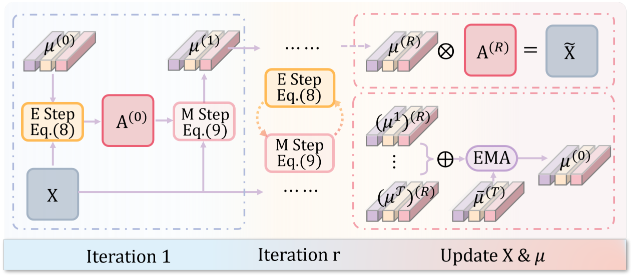

Figure 4. The structure of inter-task EM-interactive module.

cues. Finally, with the Q, $\widetilde{K}$ and $\widetilde{V}$, multi-head attention is applied to capture long-range dependency. In this way, the computational cost of our hierarchy-adaptive module is $O(CH_SW_S\sum_{s=1}^{S}\lfloor H_s/r_s\rfloor \times \lfloor W_s/r_s\rfloor)$, which is more acceptable than the naive implementation by concatenating multi-stage tokens with the quadratic computation complexity of $O(C(\sum_{s=1}^{S} H_sW_s)^2)$.

3.3. Inter-Task EM-Interactive Module

The importance of learning task-mutual information cannot be overstated, as it plays a critical role in uncovering the intricate correlations among various aspects of an input image. This process is fundamental for establishing a robust multi-task framework capable of handling the complexities and variations in different tasks. In pursuit of this objective, the most straightforward approach would be to concatenate the feature maps that are independent of each task. This concatenated features can then be utilized to compute the interdependencies across tasks by employing a multi-head self-attention (MHSA) mechanism as delineated in [41]:

$$[\widetilde{X}^1,\cdots,\widetilde{X}^{\mathcal{T}}] = \text{MHSA}([X^1,\cdots,X^{\mathcal{T}}]) \in \mathbb{R}^{C\times\mathcal{T}HW}. \quad (5)$$

However, such a naive interaction is not effective, since the built-in all-pair reliance might be confused and disturbed by inconsistent details, and thus fails to achieve the satisfactory performance. Additionally, the quadratic computation complexity of $O(C\mathcal{T}^2H^2W^2)$ may lead to the redundant calculation. Different from these methods with the concatenated tokens as $K,V \in \mathbb{R}^{C\times\mathcal{T}HW}$ in the self-attention, our inter-task interactive module (in Fig. 4) aims to find a compact set of bases $\boldsymbol{\mu} \in \mathbb{R}^{C\times K}$ from various tasks. Since $K \ll HW$, our module reduces the complexity from $O(C\mathcal{T}^2H^2W^2)$ to $O(C\mathcal{T}HWK)$, making the final pixel-wise prediction of each task more tractable. Inspired by the verified efficacy of EM [16, 29], our key idea is to find the maximum likelihood solution for exploring the latent mutual information.

Preliminaries of EM Algorithm: It is an iterative strategy for estimating parameters of models with latent variable. Given the observed data $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ and unobserved hidden variables $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_N\}$ with N samples, the goal of EM is to estimate the parameters $\theta^{(r)}$ by maximizing the likelihood in r-th iteration: $\theta^r = \arg\max \sum_{i=1}^{N} \log \sum_{\mathbf{a}_i} p(\mathbf{x}_i, \mathbf{a}_i; \theta^{(r-1)})$. In each EM iteration, two steps are involved, *i.e.,* the expectation step (E step) and the maximization step (M step):

- **E step.** It uses the posterior to find the conditional probability expectation: $Q_i(\mathbf{a}_i) = p(\mathbf{a}_i|\mathbf{x}_i, \theta^{(r-1)})$.
- **M step.** It determines the newly revised parameters by maximizing the likelihood function: $\theta^{(r)} = \arg\max \sum_{i=1}^{N}\sum_{\mathbf{a}_i} Q_i(\mathbf{a}_i) \log \frac{p(\mathbf{x}_i,\mathbf{a}_i;\theta^{(r-1)})}{Q_i(\mathbf{a}_i)}$.

Both steps are alternately executed for R iterations to converge to an optimum.

EM Algorithm for Gaussian Mixture Models (GMM): As a special case of the EM, GMM [28] models the distribution of data $\mathbf{x}_n$ as a linear superposition of K Gaussians:

- **E step.** For GMM, latent $\mathbf{a}_{nk}^{(r)}$ can be re-estimated as:

$$\mathbf{a}_{nk}^{(r)} = \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k^{(r-1)}, \mathbf{I})}{\sum_{j=1}^{K}\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j^{(r-1)}, \mathbf{I})}. \quad (6)$$

To simplify, the posterior probability of $\mathbf{x}_n$ can be formulated with the kernel function $\mathcal{K}$ as: $p(\mathbf{x}_n|\boldsymbol{\mu}_k) = \mathcal{K}(\mathbf{x}_n, \boldsymbol{\mu}_k)$. In this way, Eq. (6) can be rewritten as:

$$\mathbf{a}_{nk}^{(r)} = \frac{\mathcal{K}(\mathbf{x}_n, \boldsymbol{\mu}_k^{(r-1)})}{\sum_{j=1}^{K}\mathcal{K}(\mathbf{x}_n, \boldsymbol{\mu}_j^{(r-1)})}. \quad (7)$$

By taking the exponential inner dot $\exp\left(\mathbf{a}^\intercal\mathbf{b}\right)$ as the formulation of kernel function, Eq. (7) can be reformulated in a form similar to the attention model as:

$$\mathbf{a}_{nk}^{(r)} = \frac{\exp(\mathbf{x}_n(\boldsymbol{\mu}_k^{(r-1)})^\intercal/\tau)}{\sum_{j=1}^{K}\exp(\mathbf{x}_n(\boldsymbol{\mu}_k^{(r-1)})^\intercal/\tau)}, \quad (8)$$

where τ is the constant to adjust the distribution of $\mathbf{A}$.
- **M step.** After that, the M step is adopted to calculate the bias $\boldsymbol{\mu}^{(r)}$ with the estimated $\mathbf{A}^{(r)}$, via maximizing the likelihood. Specifically, $\boldsymbol{\mu}^{(r)}$ can be estimated by applying the weighted average on $\mathbf{X}$:

$$\boldsymbol{\mu}_k^{(r)} = \frac{1}{N_k}\mathbf{a}_{nk}^{(r)}\mathbf{x}_n, \text{ where } N_k = \textstyle\sum_{m=1}^{N}\mathbf{a}_{mk}^{(r)}. \quad (9)$$

Update X and $\boldsymbol{\mu}$ on $\mathcal{T}$ tasks. After alternating for R iterations, the parameters of GMM are converged, and the final $\boldsymbol{\mu}^{(R)}$ and $\mathbf{A}^{(R)}$ are used to reconstruct $\widetilde{\mathbf{X}}$ in Fig. 4, which can be formulated as $\widetilde{\mathbf{X}} = \mathbf{A}^{(R)}\boldsymbol{\mu}^{(R)}$.

Notably, each task-independent feature $\mathbf{X}^t$ is fed into the EM-interactive module for enhancing the representations. Specifically, they share the same initial $\boldsymbol{\mu}^{(0)}$ in the first iteration, and re-building a task-aware $(\boldsymbol{\mu}^t)^{(R)}$ (*e.g.,* the final bias for the t-th task) by the iterations. Then, we fuse the $(\boldsymbol{\mu}^t)^{(R)}$ to obtain the final $\boldsymbol{\mu}^R$ with the mutual information:

$$\boldsymbol{\mu}^{(R)} = \frac{1}{\mathcal{T}}\underbrace{((\boldsymbol{\mu}^1)^{(R)}+\cdots+(\boldsymbol{\mu}^t)^{(R)}+\cdots+(\boldsymbol{\mu}^{\mathcal{T}})^{(R)})}_{\mathcal{T}\ tasks}. \quad (10)$$

Table 1. Comparison with SOTA methods on the NYUD-V2 (left) and PASCAL-Context (right) datasets of different tasks.

NYUD-v2	Semseg	Depth	Normal	Boundary	PASCAL-Context	Semseg	Parsing	Saliency	Normal	Boundary
	mIoU ↑	RMSE ↓	mErr ↓	odsF ↑		mIoU ↑	mIoU ↑	maxF ↑	mErr ↓	odsF ↑
Cross-Stitch [26]	36.34	0.6290	20.88	76.38	PAD-Net [39]	53.60	59.60	65.80	15.30	72.50
PAD-Net [46]	36.61	0.6270	20.85	76.38	ASTMT [23]	68.00	61.10	65.70	14.70	72.40
PAP [46]	36.72	0.6178	20.82	76.42	MTI-Net [33]	61.70	60.18	84.78	14.23	70.80
PSD [48]	36.69	0.6246	20.87	76.42	ATRC [2]	62.69	59.42	84.70	14.20	70.96
MTI-Net [33]	45.97	0.5365	20.27	77.86	ATRC-ASPP [2]	63.60	60.23	83.91	14.30	70.86
ATRC [2]	46.33	0.5363	20.18	77.94	ATRC-BMTAS [2]	67.67	62.93	82.29	14.24	72.42
MQTransformer [40]	49.18	0.5785	20.81	77.00	MQTransformer [40]	71.25	60.11	84.05	14.74	71.80
DeMT [41]	51.50	0.5474	20.02	78.10	DeMT [41]	75.33	63.11	83.42	14.54	73.20
InvPT [42]	53.56	0.5183	19.04	78.10	InvPT [42]	79.03	67.61	84.81	14.15	73.00
TaskPrompter [43]	55.30	0.5152	18.47	78.20	TaskPrompter [43]	80.89	68.89	84.83	13.72	73.50
Ours	**56.82**	**0.4937**	**18.45**	**78.40**	Ours	**81.66**	**69.90**	**84.95**	**13.39**	**73.80**

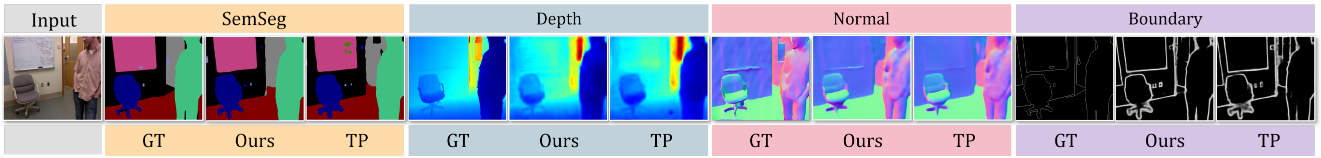

Figure 5. Visual comparisons of Ground Truth (GT), our SEM, and TaskPrompter (TP) among four tasks on the NYUD-v2 dataset.

Then, the obtained $\boldsymbol{\mu}^{(R)}$ is initialized as the $\boldsymbol{\mu}^{(0)}$ for the next batch. Notably, to avoid catastrophic forgetting from the previous stages, an exponential moving average (EMA) [16] is utilized to ensemble the information in different training steps. Hence, the update of $\boldsymbol{\mu}^{(0)}$ can be formulated as $\boldsymbol{\mu}^{(0)} = \alpha\boldsymbol{\mu}^{(0)} + (1-\alpha)\bar{\boldsymbol{\mu}}^{(T)}$, where α is the EMA decay that controls the updating rate, and $\bar{\boldsymbol{\mu}}^{(T)}$ is obtained from averaging $\boldsymbol{\mu}^{(T)}$ over a mini-batch.

Discussion. The EM-interactive design is noteworthy for its ability to reconstruct low-rank, compact task-aware features that are devoid of superfluous elements. During the iterative process, each task starts with the same initial value of $\boldsymbol{\mu}^{(0)}$, shared across all tasks. Concurrently, the representation of each task is updated to reflect its unique internal structure. As the EM iterations progress, the generated $(\boldsymbol{\mu}^{t})^{(R)}$ is reshaped and then reintegrated back into the initial $\boldsymbol{\mu}^{(0)}$ for the next input feature; taking into account the consensus of tasks, the EM algorithm effectively narrows the task disparities and enhancing task performance. Note that the learned $\boldsymbol{\mu}^{(0)}$ remains constant during the inference.

In our approach, we introduce a task-specific decoder to prevent feature corruption from other tasks. The decoder design is informed by the principle of gradually increasing spatial resolution, as suggested by [42], to preserve the essential spatial structure. Specifically, at each stage of the decoder, the input feature is first upsampled through bilinear interpolation to match the size of the corresponding encoder stage, followed by an addition operation with the encoder feature. Subsequently, the combined encoder-decoder representation is refined using a lightweight global self-attention mechanism [35]. Further details on this process can be found in the Appendix. To further enhance the decoder with interactive awareness, the refined decoder feature is input into our EM-based module, facilitating a more nuanced and effective interaction among tasks.

4. Experiments

4.1. Setup

Datasets. Experiments are conducted on two public datasets: **(1) NYUD-v2** [30] is an indoor scene dataset that pairs RGB and depth frames. It is typically used for tasks such as semantic segmentation (SemSeg), monocular depth estimation (Depth), surface normal estimation (Normal), and boundary detection (Boundary). Following the settings in [30], the dataset is divided into 795 training images and 654 testing images. **(2) PASCAL-Context** [3] is a natural scene dataset featuring 21 semantic classes. It encompasses tasks such as SemSeg, human parts segmentation (Parsing), saliency estimation (Saliency), Normal, and Boundary, with $4,998$ images for training and $5,105$ images for testing.

Metrics. We employ five metrics to compare our method with other multi-task schemes, including mean Intersection over Union (mIoU), root mean square error (RMSE), mean Error (mErr), optimal dataset scale F-measure (odsF), and maximum F-measure (maxF). Following [41, 42], Δ_m is utilized to quantify the average performance gain of multi-task models compared to single-task models.

Implementation Details. All models are trained for $40,000$ iterations with the same loss function as in [42]. Specifically, we evaluate our method on various backbones, including Swin-Transformer (Swin) and ViT (discussed in Sec. 4.4). We use the Adam optimizer with a learning rate of 1×10^{-5} for NYUD-v2 and 2×10^{-5} for PASCAL-Context, a weight decay rate of 10^{-6}, and a batch size of 4. All hyperparameter selections and ablation studies are conducted on the NYUD-v2 dataset with the ViT-L encoder.

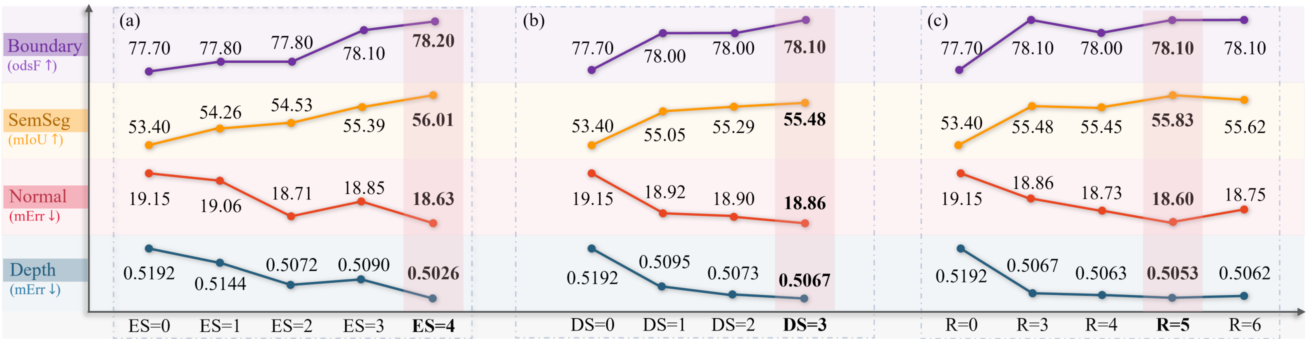

Figure 6. Hyper-parameter analysis of two task-aware modules across four tasks on the NYUD-v2 dataset: intra-task module with various encoder stages (ES) in (a); inter-task module on different decoder stages (DS) in (b) and iteration numbers (R) in (c).

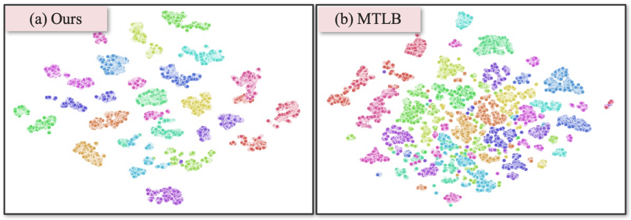

Figure 7. t-SNE visualization of deep feature representations extracted by ours (a) and the baseline (b) on PASCAL-Context.

4.2. Benchmarking Against State-of-the-Arts

We substantiate the efficacy of our SEM by conducting a series of experimental comparisons with state-of-the-arts. Notably, the two prior leading algorithms, *i,e,* InvPT [42] and TaskPrompter [43], along with our SEM, are built upon an identical ViT-L encoder for feature extraction.

Quantitative Comparison. The experimental results are highlighted in Table 1, with the top results accentuated in bold. As depicted in Table 1 (left), our approach sets a new benchmark across four metrics on NYUD-v2, outperforming the erstwhile frontrunner, TaskPrompter, by margins of 1.52% and 2.15% for semantic segmentation and depth estimation tasks, respectively. To further attest to the versatility of our method, we compare it against extant state-of-the-art models on an additional dataset (*i.e.*, PASCAL-Context). The comparative results, presented in Table 1 (right), show that our approach also achieves superior performance across five tasks, with improvements in Semseg: +0.77 in mIoU, Parsing: +1.01 in mIoU, Saliency: +0.12 in maxF, Normal: -0.33 in mErr, and Boundary: +0.30 in odsF. These gains underscore the robustness of our method in learning efficient intra-task and inter-task dependencies.

Qualitative Comparison. Fig. 5 shows visual comparisons of the top two methods on the NYUD-v2 dataset across four tasks. It is evident that our method is adept at producing high-fidelity, pixel-accurate predictions for various tasks, yielding results that are more congruent with the ground truth and exhibit enhanced clarity in detail.

4.3. Hyperparameters

Intra-task Module with Various Encoder Stages. Our initial investigation focuses on the impact of incorporating a varying number of encoder stages (ES) into our intra-task module. ViT-L encoder has 24 layers, and we treat 6 layers as a stage [42], *e.g.*, the output of layer 6, 12, 18, 24 serves as stage 1, 2, 3, 4, respectively. As shown in Fig. 6 (a), we first integrate a single stage (ES=1) from the deepest encoder, which achieves an improvement over our baseline (ES=0). By increasing the number of stages, we observe a corresponding enhancement in accuracy, with the optimal performance attained upon the inclusion of all stages.

Inter-task Module on Different Decoder Stages. Subsequently, we investigate the effects of embedding our inter-task module at various stages within the decoder (DS), with iteration $R = 3$ throughout these trials. As can be observed in Fig. 6 (b), the progressive incorporation of inter-task module across different stages yields notable performance gains, underscoring the value of facilitating full-stage interaction. Consequently, in our experiments, we incorporate the inter-task module into all decoder stages.

Inter-task Module with Iterations R. To discern the optimal setting of the iteration R, that influences the convergence of the EM-based interactions, we conduct a series of experiments across a spectrum of R values. As evidenced in Fig. 6 (c), the peak performance is attained at $R = 5$, which is adopted as the equilibrium point that harmonizes accuracy with complexity for subsequent experiments.

4.4. Ablation Study

Comparison of Multi-task and Single-task Learning. In this section, we provide a comparative analysis between our SEM and the baseline under two learning paradigms: multi-task learning (MTL) and single-task learning (STL), to evaluate the efficacy of our multi-task approach. (**i**) **MTL baseline** is anchored in ViT-L and incorporates a task-specific head characterized by a 3×3 convolution block, mirroring the head configuration of the proposed model. (**ii**) **STL**

Table 2. Ablation study of our SEM on the NYUD-v2 dataset.

#	Method	Semseg mIoU ↑	Depth RMSE ↓	Normal mErr ↓	Boundary odsF ↑	MTL Gain Δ_m ↑
1	STL Model	54.27	0.5147	18.96	77.80	-
2	MTL Model	52.39	0.5223	19.23	77.40	-1.72
3	Our baseline	53.40 (↑ **1.01**)	0.5192 (↓ **0.0031**)	19.15 (↓ **0.08**)	77.70 (↑ **0.30**)	-0.90 (↑ **0.82**)
4	Our baseline + intra-task module	56.01 (↑ **3.62**)	0.5026 (↓ **0.0197**)	18.63 (↓ **0.60**)	78.20 (↑ **0.80**)	1.95 (↑ **3.67**)
5	Our baseline + inter-task module	55.83 (↑ **3.44**)	0.5053 (↓ **0.0170**)	18.60 (↓ **0.63**)	78.10 (↑ **0.70**)	1.75 (↑ **3.47**)
6	Our baseline + inter- & inter-task	56.82 (↑ **4.43**)	0.4937 (↓ **0.0286**)	18.45 (↓ **0.78**)	78.40 (↑ **1.00**)	3.06 (↑ **4.78**)

Table 3. Compatibility of our method built upon different Transformer-based encoders on the NYUD-v2 dataset.

Encoders (for MTL)	Semseg mIoU ↑	Depth RMSE ↓	Normal mErr ↓	Boundary odsF ↑	Encoders (for ours)	Semseg mIoU ↑	Depth RMSE ↓	Normal mErr ↓	Boundary odsF ↑
Swin-T	41.71	0.6463	23.18	75.40	Swin-T	47.44	0.5669	20.43	76.50
Swin-B	49.82	0.5638	21.50	76.20	Swin-B	54.95	0.5099	19.38	77.80
ViT-B	47.23	0.5712	19.69	76.20	ViT-B	51.34	0.5222	18.95	77.60
ViT-L	52.39	0.5223	19.23	77.40	ViT-L	56.82	0.4937	18.45	78.40

baseline is architecturally identical to the MTL baseline; this model, however, is focused on one task at a time. As evidenced in Table 2, our method (#6) markedly outperforms the MTL baseline (#2), registering a cumulative gain of 4.78% across four tasks. Intriguingly, the proposed approach also demonstrates a substantial enhancement over the STL baseline (#1), with a 3.06% uptick in multi-task performance, demonstrating the capability of our model in concurrently stimulating each task.

Dissection of Intra-task and Inter-task Modules. In our ablation study, we meticulously evaluate the contributions of the intra-task module and inter-task module. Referring to Table 2, we establish our baseline (#3), which integrates a decoder (as expounded in Sec. 3.3) into the MTL baseline model (#2). By incorporating the intra-task module (#4), we observe notable improvements across all four tasks relative to our baseline. This validates the capacity of our model to harness cross-stage dependencies, thereby enhancing the internal structural learning specific to each task. In addition, the deployment of the inter-task module (#5) also elevates the performance across all tasks, yielding a multi-task gain (*i.e.,* Δ_m) of 1.75%, which is instrumental in promoting interactive learning among tasks. By synergizing the intra-task and inter-task modules, our method (#6) optimally exploits the potential of learning both task-independent and task-consensual information, mutually reinforcing each aspect to achieve peak performance through the generation of a more potent multi-task characteristic.

Integration with Different Encoders. We incorporat two distinct categories of Transformer-based encoders to determine the compatibility of our SEM. Specifically, we utilized the Swin-Transformer series (Swin-T and Swin-B) [19] and the ViT variants (ViT-B and ViT-L) [6]. The comparative results are presented in Table 3. The left columns of the table delineate the outcomes for the MTL baseline, while the right columns detail the performance metrics achieved by our SEM. The results clearly indicate that our model consistently boost the performance by a large margin across different networks and various tasks.

Distribution of Deeply Learned Features. Fig. 7 presents a comparative visualization of the deep feature representations derived from our model against those from the MTL baseline. This comparison is facilitated by the application of t-SNE on the Pascal-Context dataset. The enhanced clustering and separation indicates that our model significantly improves the discriminative capacity of the deep features, which is essential for semantic segmentation.

5. Conclusion

In this paper, we introduced synergy embedding models (SEM), a novel transformer-based architecture, which overcomes the limitations of current multi-task dense prediction methods. SEM provides an innovative intra-task module that adaptively generates salient keys/values from hierarchical encoders for an optimal trade-off, and an inter-task EM-based interaction that iteratively learns a compact set of bases from various tasks for ensuring robustness. Extensive experimental analyses validated the effectiveness of our SEM, demonstrating consistent superiority over existing state-of-the-art methods on two public benchmarks.

Acknowledgments

This work was supported in part by Zhejiang Provincial Natural Science Foundation of China (No. LZ22F020012), Major Technological Innovation Project of Hangzhou (No. 2022AIZD0147), the National Key Research and Development Project (No. 2022YFC2504605), Major Scientific Research Project of Zhejiang Lab (No. 2020ND8AD01), and Japanese Ministry for Education, Science, Culture and Sports (No. 20KK0234, No. 21H03470 and No. 20K21821).

References

[1] Deblina Bhattacharjee, Tong Zhang, Sabine Süsstrunk, and Mathieu Salzmann. Mult: An end-to-end multitask learning Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12031–12041, 2022. 1

[2] David Brüggemann, Menelaos Kanakis, Anton Obukhov, Stamatios Georgoulis, and Luc Van Gool. Exploring relational context for multi-task dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15869–15878, 2021. 6

[3] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1971–1978, 2014. 6

[4] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, pages 794–803. PMLR, 2018. 3

[5] Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. *Advances in Neural Information Processing Systems*, 33:2039–2050, 2020. 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3, 8

[7] Xiaobiao Du, Jie Niu, and Chongjin Liu. Expectation-maximization attention cross residual network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 888–896, 2021. 3

[8] Yuan Gao, Jiayi Ma, Mingbo Zhao, Wei Liu, and Alan L Yuille. NDDR-CNN: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3205–3214, 2019. 1

[9] Yuan Gao, Haoping Bai, Zequn Jie, Jiayi Ma, Kui Jia, and Wei Liu. MTL-NAS: Task-agnostic neural architecture search towards general-purpose multi-task learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 11543–11552, 2020. 1

[10] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019. 1

[11] Jianzhong He, Shiliang Zhang, Ming Yang, Yanhu Shan, and Tiejun Huang. Bi-directional cascade network for perceptual edge detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3837, 2019. 1

[12] Varun Jampani, Deqing Sun, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Superpixel sampling networks. In *Proceedings of the European Conference on Computer Vision*, pages 352–368, 2018. 3

[13] Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-language representations. *Advances in Neural Information Processing Systems*, 35:30291–30306, 2022. 3

[14] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7482–7491, 2018. 1, 3

[15] Bo Li, Chunhua Shen, Yuchao Dai, Anton van den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1119–1127, 2015. 1

[16] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9167–9176, 2019. 2, 3, 5, 6

[17] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-task multi-sensor fusion for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7345–7353, 2019. 1

[18] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1871–1880, 2019. 1

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision Transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1, 3, 8

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 1

[21] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Wei Li, Christian Theobalt, Ruigang Yang, and Wenping Wang. Adaptive surface normal constraint for depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12849–12858, 2021. 1

[22] Ivan Lopes, Tuan-Hung Vu, and Raoul de Charette. Cross-task attention mechanism for dense multi-task learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2329–2338, 2023. 1

[23] Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. Attentive single-tasking of multiple tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1851–1860, 2019. 6

[24] Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. Attentive single-tasking of multiple tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1851–1860, 2019. 1

[25] Etienne Meunier, Anaïs Badoual, and Patrick Bouthemy. EM-driven unsupervised learning for efficient motion segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4462–4473, 2022. 2

[26] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003, 2016. 6

[27] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. BASNet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7479–7489, 2019. 1

[28] Sylvia. Richardson and Peter J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, 59(4):731–792, 2002. 5

[29] Nima Sammaknejad, Yujia Zhao, and Biao Huang. A review of the expectation maximization algorithm in data-driven process identification. *Journal of Process Control*, 73:123–136, 2019. 2, 5

[30] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Proceedings of the European Conference on Computer Vision*, pages 746–760. Springer, 2012. 6

[31] Chonghyuk Song, Eunseok Kim, and Inwook Shim. Improving gradient flow with unrolled highway expectation maximization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9704–9712, 2021. 3

[32] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 1

[33] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. MTI-Net: Multi-scale task interaction networks for multi-task learning. In *Proceedings of the European Conference on Computer Vision*, pages 527–543. Springer, 2020. 1, 6

[34] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3614–3633, 2021. 1, 2

[35] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision Transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 6

[36] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CVT: Introducing convolutions to vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 1

[37] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision Transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4794–4803, 2022. 4

[38] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with Transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090, 2021. 1

[39] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. 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*, pages 675–684, 2018. 1, 6

[40] Yangyang Xu, Xiangtai Li, Haobo Yuan, Yibo Yang, and Lefei Zhang. Multi-task learning with multi-query Transformer for dense prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1, 3, 6

[41] Yangyang Xu, Yibo Yang, and Lefei Zhang. DeMT: Deformable mixer Transformer for multi-task learning of dense prediction. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 3072–3080, 2023. 1, 2, 3, 5, 6

[42] Hanrong Ye and Dan Xu. Inverted pyramid multi-task Transformer for dense scene understanding. In *Proceedings of the European Conference on Computer Vision*, pages 514–530. Springer, 2022. 1, 2, 3, 6, 7

[43] Hanrong Ye and Dan Xu. TaskPrompter: Spatial-channel multi-task prompting for dense scene understanding. In *International Conference on Learning Representations*, 2023. 1, 3, 6, 7

[44] Luhui Yue, Junxia Li, and Qingshan Liu. Body parts relevance learning via expectation–maximization for human pose estimation. *Multimedia Systems*, 27:927 – 939, 2021. 2

[45] Jianxin Zhang, Zongkang Jiang, Dongwei Liu, Qiule Sun, Yaqing Hou, and Bin Liu. 3D asymmetric expectation-maximization attention network for brain tumor segmentation. *NMR in Biomedicine*, 35(5):e4657, 2022. 3

[46] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4106–4115, 2019. 6

[47] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6881–6890, 2021. 1

[48] Ling Zhou, Zhen Cui, Chunyan Xu, Zhenyu Zhang, Chaoqun Wang, Tong Zhang, and Jian Yang. Pattern-structure diffusion for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4514–4523, 2020. 6