

Efficient Computation Sharing for Multi-Task Visual Scene Understanding

Sara Shoouri

Mingyu Yang

Zichen Fan

Hun-Seok Kim

University of Michigan

{sshouri, mingyuy, zcfan, hunseok}@umich.edu

Abstract

Solving multiple visual tasks using individual models can be resource-intensive, while multi-task learning can conserve resources by sharing knowledge across different tasks. Despite the benefits of multi-task learning, such techniques can struggle with balancing the loss for each task, leading to potential performance degradation. We present a novel computation- and parameter-sharing framework that balances efficiency and accuracy to perform multiple visual tasks utilizing individually-trained single-task transformers. Our method is motivated by transfer learning schemes to reduce computational and parameter storage costs while maintaining the desired performance. Our approach involves splitting the tasks into a base task and the other sub-tasks, and sharing a significant portion of activations and parameters/weights between the base and sub-tasks to decrease inter-task redundancies and enhance knowledge sharing. The evaluation conducted on NYUD-v2 and PASCAL-context datasets shows that our method is superior to the state-of-the-art transformer-based multi-task learning techniques with higher accuracy and reduced computational resources. Moreover, our method is extended to video stream inputs, further reducing computational costs by efficiently sharing information across the temporal domain as well as the task domain. Our codes are available at <https://github.com/sarashoouri/EfficientMTL>.

1. Introduction

In various computer vision applications, it is required to obtain a comprehensive understanding of the visual scene by performing multiple tasks based on a single input image. These tasks often involve performing dense or pixel-wise predictions such as semantic segmentation, depth estimation, surface normal estimation, etc., for practical applications in autonomous driving, robotics, and augmented or virtual reality (AR/VR) [68]. Traditionally, these tasks are tackled individually by training a separate neural network dedicated to each task. However, this single-task learning can lead

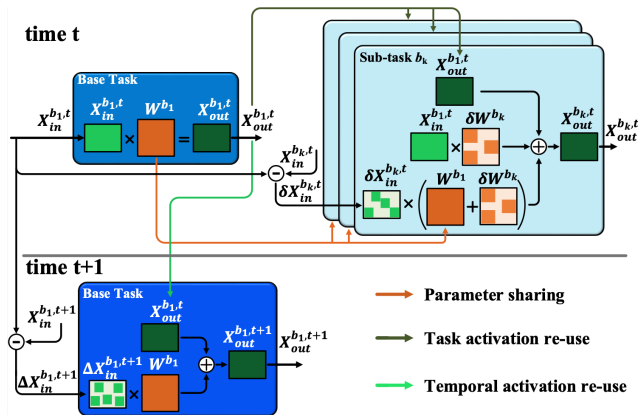


Figure 1: Parameter and activation sharing across task and temporal domains. First, the base task produces its activations at time t , then passes them to sub-tasks b_k to share the computation across the task domain. Also, the base task activations at time t are shared with time $t + 1$ to reduce the computation across the temporal domain.

to redundant computation and parameters, particularly for highly correlated tasks, losing the opportunity to perform faster inference as desired in real-time applications [59].

Multi-task learning (MTL) [6, 16, 34, 70, 73] has been actively explored as a solution to this problem, which learns a single unified model to perform multiple tasks simultaneously. This approach allows the model to learn common representations and patterns from multiple supervised tasks [74], resulting in a more memory- and computation-efficient structure than employing multiple single-task networks. Additionally, multi-task networks can potentially enhance the performance of all tasks by leveraging the cross-task knowledge-sharing mechanism [59]. Due to the benefits of multi-task learning, numerous multi-task structures have been proposed to improve task-wise interactions in dense visual scene understanding [43, 65, 59, 75, 77, 42, 18]. The recent success of transformer models in many downstream vision tasks [37, 13, 78, 61, 49, 38] has led to the emergence of transformer-based multi-task networks [51, 2, 68, 7, 44], aimed at improving multi-task performance even further.

However, MTL poses several potential drawbacks compared to single-task learning: (1) Simultaneously learning different tasks using a unified model can lead to unbalanced task competition and suboptimal performance for some tasks if the model fails to build a shared representation that generalizes to all tasks [17]. (2) Balancing the loss between different tasks can be challenging due to the varying scales of task-specific loss terms, especially when the number of tasks increases [11, 9]. Although multiple task-balancing approaches have been proposed [30, 9, 35] to address this problem, recent studies [60] have shown that performance still becomes worse than individually-trained single-task networks. (3) MTL requires ground truth labels for all tasks per training sample, which is limiting because such annotations for certain tasks (e.g., semantic segmentation) may not always be available for the same input sample. It also often does not allow easily adding new tasks without retraining the entire model from scratch.

To alleviate the above limitations, we propose a solution that leverages the strength of both single-task and multi-task learning techniques, integrating the concept of knowledge-sharing. Our approach draws inspiration from [10], which initially introduced the idea of efficient deep learning model communication through knowledge-sharing. In our work, we extend this concept to effectively execute multiple concurrent visual tasks taking the same input. We employ individually-trained single-task networks to maintain the desired performance and prevent unbalanced task competition. At the same time, we introduce a novel parameter- and computation-sharing strategy to facilitate knowledge-sharing and enhance inference efficiency. Our method focuses on transformer models that have shown outstanding results in vision tasks. First, we divide all tasks into a base task and multiple sub-tasks, where all task-specific networks adopt a common transformer structure as the *backbone*. Next, we train a single network for the base task independently. Then, to share the inter-task information, we reuse *not only weights but also activations* from the base task to train each sub-task. The weight-sharing concept is motivated by recent parameter-efficient transfer learning techniques [26, 21]. Specifically, we view the weights of sub-task models as the sum of weights from the base task and a *delta weight* matrix, and apply ℓ_0 regularization to encourage sparsity in the delta weight matrix as in the Diff-pruning [21] approach. We then fix the positions of non-zero elements of the delta weight matrix and fine-tune them to make the *activation difference* between the base task and sub-task also sparse by adopting ℓ_1 regularization. As a result, the pre-computed activations from the base task can be shared and passed to sub-task networks during the inference. This reduces computation cost as the remaining operations for sub-tasks only involve sparse matrix-matrix multiplications, as shown in Figure 1. Our proposed computation-sharing scheme significantly

reduces the number of non-zero parameters for sub-tasks while allowing knowledge sharing between the main task and each sub-task. Extensive experiments on NYUD-v2 and Pascal-Context benchmarks demonstrate that our method outperforms state-of-the-art multi-task transformers, exhibiting fewer parameters and FLOPs counts to attain comparable or superior task accuracy.

Furthermore, we extend our method to the temporal domain to leverage the sparsity of differences between consecutive video frames, as shown in Figure 1. Similar to the task domain, we employ ℓ_1 regularization to force the activation differences between consecutive (time domain) frames to become sparse. As a result, during the inference of a sub-task at time t , it can exploit either the temporal domain or task domain sparsity to reuse activations from the same sub-task at time $t - 1$ (temporal activation sharing) or the main task at time t (task domain activation sharing). A simple strategy is then applied to combine these two sources of activation sharing for maximum efficiency and computation savings.

Our contributions can be summarized as follows:

- We propose a novel activation-and parameter-sharing scheme to reduce the computation and storage redundancy to perform multiple dense/pixel-wise vision tasks with transformer models.
- We extend our computation-sharing method to video inputs and use a simple strategy to combine the activation-sharing sources between the task and temporal domains to maximize inference efficiency.
- We perform extensive experiments on NYUD-v2 and Pascal-Context datasets to quantify the advantage of our proposed method in both performance and efficiency compared to prior multi-task learning methods.

2. Related Work

2.1. Multi-task for Scene Understanding

Multi-task learning (MTL) has made significant progress in several vision applications [6, 16, 32, 30], such as joint object detection and semantic segmentation [19, 23], and 3D recognition [66, 67]. Due to the benefits of MTL, various works have explored its use in multi-task dense scene understanding, where all the tasks require pixel-wise predictions [15, 43] based on convolutional neural networks (CNN) [18, 3, 42]. Specifically, Cross-Stitch [43] enables soft feature fusion by utilizing a linear combination of the activations in each layer of task-specific networks. PAD-Net [65] first generates multi-modal auxiliary predictions and then uses a spatial attention strategy to distill information from the initial predictions. MTI-Net [59] extends the idea of PAD-Net by proposing a multi-scale multi-modal distillation procedure to refine the feature propagation, leading to unique task interactions at each individual scale. On the other hand, PAP [75] and PSD [77] learn global and local

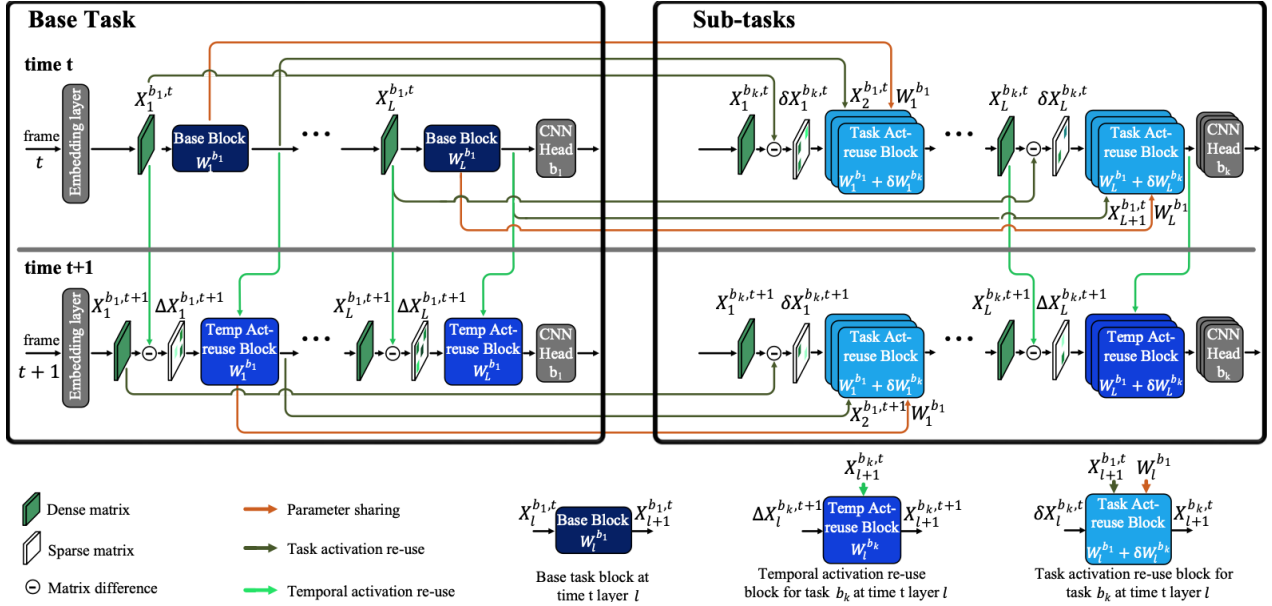


Figure 2: Illustration of the proposed method for multi-task processing for video inputs. $X_l^{b_k,t}$, $\delta X_l^{b_k,t}$, and $\Delta X_l^{b_k,t}$ represent the input activation, delta task activation, and delta temporal activation of layer l for task b_k at time t , respectively. For a keyframe at t , the base task shares the parameter and activation computation with sub-tasks to improve the efficiency of the sub-task transformer models. For a non-keyframe at $t + 1$, the base task first borrows the temporal activation from the previous frame at t and then passes the calculated task activation to sub-tasks for layer $[1, l]$ at $t + 1$. For the remaining layers of $[l + 1, L]$ at $t + 1$, the sub-tasks reuse the temporal activation from the previous time t to further reduce the computation.

affinities across tasks, and use them to refine the features for each task iteratively. Recently, the transformer model has also been applied to multi-task learning [7, 44, 51, 2], replacing CNNs for superior performance in several vision tasks [37, 13, 1, 4, 46, 48, 57, 62]. For instance, InvPT [68] leverages vision transformers as the backbone and explores self-attention for better spatial and task interaction, achieving state-of-the-art results for multi-task visual scene understanding.

Our approach differs from prior multi-task learning methods which mainly focus on enhancing cross-task interaction. Specifically, we utilize individually-trained single-task networks with a novel computation-sharing scheme to improve inference efficiency while preserving their performances to outperform prior multi-task learning methods.

2.2. Parameter-Efficient Transfer learning

Transfer learning and multi-task learning are two complementary machine learning paradigms that share a common goal of improving model efficiency through knowledge transfer. Transfer learning can be viewed as a specialized case of multi-task learning, where the knowledge (i.e., extracted features) is transferred from a source task or domain to improve performance on a related target task. This technique has been widely adopted in various applications, including computer vision [5, 45, 54], natural language processing (NLP)

[20, 28, 76, 50, 36, 53], and speech recognition [56, 55]. The most common transfer learning methods for transformer-based models involve training a shared model, and then either fine-tuning model parameters or using linear probes with additional multi-layer perceptron (MLP) models [64, 58, 14] for individual tasks. The fine-tuning method can maintain high performance at the cost of high computational complexity for large transformers [47], whereas linear probing exhibits relatively low performance with reduced complexity [33, 31, 72, 22]. Recently, there have been several research efforts focused on improving the parameter efficiency for transformers in NLP [26, 21, 27, 69, 29]. For example, Adapter [26] addresses this challenge by employing additional task-specific trainable modules after the feedforward network in each layer of the shared transformer architecture. Diff-pruning [21] further improves parameter efficiency by learning sparse task-specific difference (‘diff’ or ‘delta’) vectors to be added to the shared parameters. LoRA [27] takes a similar approach but decomposes the task-specific difference vector as a low-rank matrix. Parameter-efficient transfer learning has also been studied for vision transformers [24]. Unlike these prior methods that mainly focus on improving parameter efficiency, our approach extends the existing methods to improve both parameter and computation efficiency when multiple pixel-level tasks are performed on a single still image or consecutive video frames.

3. Proposed Method

The proposed approach is designed to enable the efficient processing of multiple dense (per pixel) tasks from a single input image. The approach is further optimized for video sequences to balance accuracy and efficiency, reducing computational costs by exploiting either the task or temporal domain sparsity as depicted in Figure 2. We first formalize the problem for the algorithm and then present a general framework that leverages activation reuse across the task or temporal domain.

3.1. Framework Overview

Problem formulation: The proposed multi-task model, \mathcal{G} , is designed to perform a set of m visual tasks, $\mathcal{B} = \{b_1, b_2, \dots, b_m\}$, on a single input image I . It generates the task outputs, $\hat{y} = \{\hat{y}^{b_1}, \hat{y}^{b_2}, \dots, \hat{y}^{b_m}\}$, for each task in \mathcal{B} . Here we assume per-pixel tasks such as semantic segmentation or depth estimation. The model can be further applied to a set of video frames $\{I^1, \dots, I^T\}$, where I^t is the input frame at time t . Hence, the model output for the frame at time t can be represented as $\mathcal{G}(I^t) \rightarrow \hat{y}^t = \{\hat{y}^{b_1,t}, \hat{y}^{b_2,t}, \dots, \hat{y}^{b_m,t}\}$.

We begin by dividing the desired tasks into a base task and the other sub-tasks. Without loss of generality, suppose b_1 is the base task and the remaining tasks, $\{b_2, \dots, b_m\}$, are the sub-tasks. To perform each task, we define separate single-task networks and leverage transfer learning techniques to initialize the parameters of the sub-tasks using the base model and then fine-tune each model for a task-specific objective.

Suppose a training dataset $\mathcal{D}^{b_k} = \{I_n, y_n\}_{n=1}^N$ is given for a single network for task b_k . We seek to determine the optimal parameters w^{b_k} for that model by solving the following optimization problem:

$$\min_{w^{b_k}} \frac{1}{N} \sum_{n=1}^N \mathcal{C}^{b_k}(f^{b_k}(I_n; w^{b_k}), y_n) + \lambda \mathcal{R}(w^{b_k}), \quad (1)$$

where $f^{b_k}(\cdot; w^{b_k})$ represents a task-specific network, $\mathcal{C}^{b_k}(\cdot)$ is a task-specific loss function, and $\mathcal{R}(\cdot)$ is an optional regularization term with a hyperparameter λ . For the network f^{b_k} , we use a vision transformer [13, 37] as the backbone, E^{b_k} , followed by a small task-specific CNN head, H^{b_k} . That is, $f^{b_k}(\cdot) = H^{b_k}(E^{b_k}(\cdot))$ holds. While it is beneficial to have a separate model for each sub-task, such an approach can be inefficient due to the high memory and computation requirements during both training and inference, making it impractical for resource-constrained real-world applications. To address this issue, our method applies delta weight, delta temporal activation, and delta task activation pruning techniques to the E^{b_k} for all existing sub-tasks as described in Sec. 3.2 and 3.3. The CNN head H^{b_k} for each sub-task is trained without exploiting delta weight/activation sparsity.

Transformer backbone E^{b_k} : The transformer model consists of two fundamental computation elements in each layer: a multi-head self-attention module and a feedforward network (FFN). The multi-head self-attention module calculates the inter-dependencies between different positions in the input data through the following equations:

$$\begin{aligned} MultiHead(Q, K, V) &= Concat(head_1, \dots, head_h)W_O \\ \text{where } head_i &= Attention(XW_Q, XW_K, XW_V), \quad (2) \\ Attention(Q, K, V) &= softmax(QK^T/\sqrt{D})V, \end{aligned}$$

where $X \in R^{P \times D}$ is the activation input that has P entries of D dimension, $\{W_Q, W_K, W_V\} \in R^{D \times D}$, and $W_O \in R^{Dh \times D}$ are model parameters, and h is the number of heads. An FFN is applied to the multi-head output of the self-attention layers. It contains two linear projections $\{W_{F1}, W_{F2}\} \in R^{D \times D}$, connected by a GELU [25] non-linearity σ , and it can be expressed as $FFN(X) = \sigma(XW_{F1})W_{F2}$. Therefore, each transformer block contains $4h + 2$ linear projections, each with a computational complexity of $O(PD^2)$. On the other hand, the calculation of the self-attention module, represented as $Attention(\cdot)$, has a complexity of $O(P^2D)$. As such, reducing the computations for linear projections can be equally important as reducing the calculation of the self-attention module. The objective of our technique is to reduce the complexity of linear projections.

Framework: For simplicity, we describe the proposed activation reuse method for an arbitrary linear projection of a vision transformer block. Given the input activation $X_{in}^{b_1,t}$ and weight matrix W^{b_1} for the base task b_1 at time t , the model first generates the outputs without activation reuse such that $X_{out}^{b_1,t} = X_{in}^{b_1,t} W^{b_1}$. For sub-task b_k , the model first learns a sparse task-specific delta weight matrix δW^{b_k} and forms the final task-specific weight as $W^{b_k} = W^{b_1} + \delta W^{b_k}$, reusing the weights from the main task. This approach reduces the storage requirements of each sub-task by only storing the sparse δW^{b_k} instead of the dense matrix W^{b_k} . However, the activation computation for b_k involves $X_{out}^{b_k,t} = X_{in}^{b_k,t} (W^{b_1} + \delta W^{b_k})$, which has the same (or higher due to weight additions) complexity as the base task. To reduce the complexity, the computation is reformulated by introducing a delta task activation matrix, $\delta X_{in}^{b_k,t} = X_{in}^{b_k,t} - X_{in}^{b_1,t}$ with a goal to remove inter-task redundancies by making $\delta X_{in}^{b_k,t}$ sparse. This allows for activation sharing from the base task b_1 to sub-task b_k , resulting in a more efficient calculation, which can be written as:

$$\begin{aligned} X_{out}^{b_k,t} &= (X_{in}^{b_1,t} + \delta X_{in}^{b_k,t}) (W^{b_1} + \delta W^{b_k}) \\ &= X_{in}^{b_1,t} W^{b_1} + \delta X_{in}^{b_k,t} (W^{b_1} + \delta W^{b_k}) + X_{in}^{b_1,t} \delta W^{b_k}. \quad (3) \end{aligned}$$

As expressed in Eq. (3), our method borrows the activation computation from the base task $X_{in}^{b_1,t} W^{b_1}$ for sub-tasks

and executes only sparse matrix-matrix multiplications in Eq. (3) rather than performing a dense matrix-matrix multiplication $X_{in}^{b_k,t} W^{b_k}$. Thus, the new calculation of a sub-task is reduced to $X_{Re}^{b_k,t} = \delta X_{in}^{b_k,t} (W^{b_1} + \delta W^{b_k}) + X_{in}^{b_1,t} \delta W^{b_k}$, where δW^{b_k} and $\delta X_{in}^{b_k,t}$ are sparse matrices. This not only increases the speed of executing the sub-tasks but also reduces the number of required (non-zero) parameters. Additionally, to ensure that the $X_{Re}^{b_k,t}$ remains sparse for the next layer, we set a per-block task-specific threshold th^{b_k} such that the activations whose absolute values are less than the predefined threshold th^{b_k} are set to zero before being passed to the next layer. The final output activation can be written as:

$$X_{out}^{b_k,t} = X_{out}^{b_1,t} + Q(X_{Re}^{b_k,t}, th^{b_k}), \quad (4)$$

where $Q(X, th)$ is an operator to substitute all elements less than th in X with zero.

Eq. (3) and (4) enable the activation sharing across different task domains (from b_1 to b_k). The same activation-sharing idea can also be applied to the temporal domain so that activations are shared from time t to time $t+1$ for the same task in \mathcal{B} . Given the input activation $X_{in}^{b_k,t}$ and weight matrix W^{b_k} for an arbitrary task b_k , we first perform the linear projection at time t to obtain $X_{out}^{b_k,t} = X_{in}^{b_k,t} W^{b_k}$. To reduce the complexity for the next time step $t+1$, we introduce a delta temporal activation matrix $\Delta X_{in}^{b_k,t+1} = X_{in}^{b_k,t+1} - X_{in}^{b_k,t}$ and exploit temporal redundancies by making $\Delta X_{in}^{b_k,t+1}$ sparse. Hence, the calculation of task b_k at time $t+1$ can be optimized by borrowing activations from time t , as follows:

$$\begin{aligned} X_{out}^{b_k,t+1} &= X_{in}^{b_k,t+1} W^{b_k} = (X_{in}^{b_k,t} + \Delta X_{in}^{b_k,t+1}) W^{b_k} \\ &= X_{in}^{b_k,t} W^{b_k} + \Delta X_{in}^{b_k,t+1} W^{b_k}. \end{aligned} \quad (5)$$

Similar to the task domain, we can remove the dense matrix-matrix multiplication $X_{in}^{b_k,t+1} W^{b_k}$ by reusing the activations from the previous time step t , and only process $\Delta X_{in}^{b_k,t+1} W^{b_k}$ which is a sparse matrix-matrix multiplication, as in Eq. (5).

In the following sections, we will describe our strategy to prune the delta weight, delta task activation, and delta temporal activation to be as sparse as possible. Moreover, we will explain how the task- and temporal-domain activation sharings are combined to minimize the task complexity.

3.2. Learning and pruning delta weight

To simplify, we consider the collection of all linear layers in a transformer block as a set, denoted by $\Phi = \{W_Q^1, W_K^1, W_V^1, \dots, W_Q^h, W_K^h, W_V^h, W_O, W_{F1}, W_{F2}\}$. Given the weights of the sub-task b_k and the base task b_1 , we aim to learn a sparse delta weight matrix that can

*We use ΔX and δX for matrix differences across temporal and task domains, respectively.

reparameterize the task-specific model as $\Phi^{b_k} = \Phi^{b_1} + \delta\Phi^{b_k}$. In this reparameterization, the weight of the base task Φ^{b_1} remains fixed during the fine-tuning process, only updating $\delta\Phi^{b_k}$ for the target sub-task. To promote sparsity in $\delta\Phi^{b_k}$, we follow the approach of Diff-pruning [21] and apply ℓ_0 regularization to $\delta\Phi^{b_k}$. Hence, the optimization loss function in Eq. (1) is modified to:

$$\min_{\delta\Phi^{b_k}} \frac{1}{N} \sum_{n=1}^N \mathcal{C}^{b_k}(f^{b_k}(I_n; \Phi^{b_1} + \delta\Phi^{b_k}), y_n) + \mathcal{R}_w(\delta\Phi^{b_k}), \quad (6)$$

where $\mathcal{R}_w(\delta\Phi^{b_k})$ is defined such that:

$$\mathcal{R}_w(\delta\Phi^{b_k}) = \lambda_w \|\text{vec}(\delta\Phi^{b_k})\|_0 = \lambda_w \sum_j \mathbb{1}\{\delta\Phi_j^{b_k} \neq 0\}, \quad (7)$$

where $\delta\Phi_j^{b_k}$ is the j^{th} element of $\delta\Phi^{b_k}$. Since this regularization term is non-differentiable, we adopt a gradient-based learning approach that uses a relaxed binary mask matrix as described in [21, 40]. This involves defining a binary mask M^{b_k} for sub-task b_k and relaxing it into a continuous space using a stretched Hard-Concrete distribution [63, 41], allowing a differentiable gradient path. The resulting mask is then element-wise multiplied with the dense delta weight matrix $\delta\Phi^{b_k}$ to produce a sparsified version.

3.3. Pruning delta task and temporal activations

Delta task and temporal activation: Similarly, we define \mathcal{X} as a set of the activation inputs of all linear layers in a transformer block. Hence, the activation input difference between the base task b_1 and the sub-task b_k at time t is represented by $\delta\mathcal{X}^{b_k,t}$. Our goal is to prune these activation differences to minimize inter-task redundancies, as discussed in Sec. 3.1. To encourage the sparsity in $\delta\mathcal{X}^{b_k,t}$, we fix the position of non-zero elements of the delta weight $\delta\Phi^{b_k}$, and then fine-tune non-zero delta weights by applying regularization to the activation differences. As $\delta\mathcal{X}^{b_k,t}$ depends on both the input image and task-specific weight matrix, it is challenging to follow the same approach used in delta weight pruning and learn a fixed binary mask which applies to all inputs. Thus, as an alternative approach, we use ℓ_1 regularization (also known as Lasso), which utilizes a Laplacian-like distribution to increase the amounts of small values. The delta task activation regularization \mathcal{R}_{a1} with a coefficient of λ_{a1} is defined as follows:

$$\mathcal{R}_{a1}(\delta\mathcal{X}^{b_k,t}) = \lambda_{a1} \|\text{vec}(\delta\mathcal{X}^{b_k,t})\|_1 = \lambda_{a1} \sum_j |\delta\mathcal{X}_j^{b_k,t}|, \quad (8)$$

where $\delta\mathcal{X}_j^{b_k,t}$ is j^{th} element of $\delta\mathcal{X}^{b_k,t}$.

The same technique extends to the temporal redundancy pruning (Sec. 3.1) to sparsify the delta temporal activation within the same (sub-)task. Consider $\Delta\mathcal{X}^{b_k,t+\tau}$ as the activation input difference between time t and $t+\tau$ for task b_k .

Table 1: Performance comparison on PASCAL-Context (left) and on NYUD-v2 (right). ‘↑’: higher better, ‘↓’: lower better.

Method	Semseg	Parsing	Saliency	Normal	Boundary	Method	Semseg	Depth	Normal	Boundary
	mIoU ↑	mIoU ↑	maxF ↑	mErr ↓	odsF ↑		mIoU ↑	RMSE ↓	mErr ↓	odsF ↑
ASTMT [42]	68.00	61.10	65.70	14.70	72.40	Cross-Stitch [43]	36.34	629.0	20.88	76.38
PAD-Net [65]	53.60	59.60	65.80	15.30	72.50	PAP [75]	36.72	617.8	20.82	76.42
MTI-Net [59]	61.70	60.18	84.78	14.23	70.80	PSD [77]	36.69	624.6	20.87	76.42
ATRC [3]	62.69	59.42	84.70	14.20	70.96	PAD-Net [65]	36.61	627.0	20.85	76.38
ATRC-ASPP [3]	63.60	60.23	83.91	14.30	70.86	MTI-Net [59]	45.97	536.5	20.27	77.86
ATRC-BMTAS [3]	67.67	62.93	82.29	14.24	72.42	ATRC [3]	46.33	536.3	20.18	77.94
InvPT (ViT-B) [68]	77.33	66.62	85.14	13.78	73.20	InvPT (ViT-B) [68]	50.30	536.7	19.00	77.60
Ours (ViT-B)	78.24	71.71	85.28	13.65	78.6	Ours (ViT-B)	50.46	533.2	18.42	77.89

Table 2: Performance and complexity of our proposed method with different levels of sharing on (a) PASCAL-Context dataset and (b) NYUD-v2 for a single image.

(a) Performance results for Pascal-Context dataset.

Method	Semseg	Parsing	Saliency	Normal	Boundary	Param (M)	Memory (Mb)	Total	Base-task	Per Sub-task
	mIoU ↑	mIoU ↑	maxF ↑	mErr ↓	odsF ↑			FLOPs (G)	FLOPs (G)	FLOPs (G)
Single Task (ST)	78.24	73.85	85.33	13.32	79.9	438.02	1776	585.15	120.53	116.15
InvPT	77.33	66.62	85.14	13.78	73.2	176.35	705	412.17	82.43	82.43
Ours + weight sharing	78.24	73.16	85.42	13.34	79.9	114.02	482	585.15	120.53	116.15
Ours + weight + act sharing	78.24	71.71	85.28	13.65	78.6	114.02	482	295.84	120.53	43.76

(b) Performance results for NYUD-v2 dataset.

Method	Semseg	Depth	Normal	Boundary	Param (M)	Memory (Mb)	Total	Base-task	Per Sub-task
	mIoU ↑	RMSE ↓	mErr ↓	odsF ↑			FLOPs (G)	FLOPs (G)	FLOPs (G)
Single Task (ST)	50.48	521.5	17.26	77.84	366.64	1452	314.14	75.20	79.64
InvPT	50.30	536.7	19.00	77.6	160.57	643	229.43	57.35	57.35
Ours + weight sharing	50.46	529.6	17.62	77.84	122.46	477	314.14	75.20	79.64
Ours + weight + act sharing	50.46	533.2	18.42	77.89	122.46	477	186.88	75.20	37.22

To encourage sparsity in $\Delta\mathcal{X}^{b_k, t+\tau}$, we define the delta temporal activation regularization \mathcal{R}_{a2} with a coefficient of λ_{a2} such that $\mathcal{R}_{a2}(\Delta\mathcal{X}^{b_k, t+\tau}) = \lambda_{a2} \|vec(\Delta\mathcal{X}^{b_k, t+\tau})\|_1$. We apply temporal regularization to $\tau \in [-2, 2]$ and reformulate the optimization problem accordingly:

$$\min_{\delta\Phi^{b_k}} \frac{1}{NT} \sum_{t=1}^T \sum_{n=1}^N C^{b_k}(f^{b_k}(I_n^t; \Phi^{b_1} + M^{b_k} \delta\Phi^{b_k}), y_n^t) + \mathcal{R}_{a1}(\delta\mathcal{X}^{b_k, t}) + \sum_{\tau=-2}^2 \mathcal{R}_{a2}(\Delta\mathcal{X}^{b_k, t+\tau}). \quad (9)$$

Activation re-use combination from both domains: To efficiently combine activations from both the task and temporal domains, we compare the computational complexities of two approaches per layer. While the base task can only reuse activations across the temporal domain, sub-tasks can access activations from both sources: using either $\delta X_{in}^{b_k, t}$ as in Eq. (3) or using $\Delta X_{in}^{b_k, t+1}$ as in Eq. (5). Suppose the (average) density (ratio of non-zero values) of the delta weight, delta task activation, and delta temporal activation for layer l is denoted by S_w^l , S_{a1}^l , and S_{a2}^l , respectively. Then, Eq. (3) requires $(S_w^l + S_{a1}^l)PD^2$ multiplications, while Eq. (5) requires $(S_{a2}^l)PD^2$ multiplications. Thus, if $S_w^l + S_{a1}^l < S_{a2}^l$, it is more efficient to reuse activations across the task domain. Conversely, if $S_{a2}^l \leq S_w^l + S_{a1}^l$, it is more reasonable to reuse activations from the previous frame of the same

(sub-)task. We observed that the first few layers tend to be more sparse in the task domain, while the remaining layers are more sparse in the temporal domain. After evaluating S_w^l , S_{a1}^l , and S_{a2}^l for all layers in each sub-task b_k , we determine the layer boundary l^{b_k} so that the task domain activation reuse is performed for layers $l \leq l^{b_k}$, and temporal domain activation reuse is utilized for the remaining layers.

4. Experiments

Dataset: We evaluate our algorithm on two popular scene understanding datasets, NYUD-v2 [52] and PASCAL-Context [8]. NYUD-v2 contains 1,449 indoor scene images with annotations for semantic segmentation (40 classes), depth estimation, surface normal estimation, and edge detection tasks, including 795 images for training and 654 for testing. PASCAL-Context covers indoor and outdoor scenes and comprises 4,998 training and 5,105 testing images, providing the labels for human parsing, semantic segmentation (21 classes), saliency estimation, edge detection, and surface normal estimation. For video input, we limit our analysis to the NYUD-v2 dataset, as PASCAL-Context does not provide images for different time frames.

Evaluation Metrics: As in InvPT [68], semantic segmentation and human parsing are evaluated with mean Intersection over Union (mIoU), surface normal estimation with mean error (mErr), depth estimation with root mean square

error (RMSE) in millimeter, edge detection with optimal-dataset-scale F-measure (odsF), and saliency detection with maximum F_1 score (maxF).

Training Details: We perform our experiments using the ViT-B transformer [13] pre-trained on ImageNet-22K [12] as the backbone, with a patch size of 16×16 pixels. Training on the NYUD-v2 dataset is conducted using a batch size of 64, distributed across 5 NVIDIA A40 single-precision GPUs, taking approximately 24 hours. The AdamW optimizer [39] is utilized, with a learning rate of 1×10^{-4} and a weight decay rate of 1×10^{-6} . For the PASCAL-Context dataset, a batch size of 6, a learning rate of 5×10^{-5} , and a weight decay rate of 1×10^{-6} are used. Training on this dataset is performed on 6 NVIDIA A40 single-precision GPUs, taking around 40 hours. Both datasets are trained using a polynomial learning rate scheduler [71].

Our approach considers ‘semantic segmentation’ as the base task and all other tasks as sub-tasks. To train these tasks, we first train the base task and store the weights and intermediate activations as the base weights and activations. Then, we follow a three-step training procedure for the sub-tasks. First, the sparse delta weight matrices are learned for each sub-task with ℓ_0 regularization on delta weights. Then, the model is fine-tuned for a few epochs, and the learned delta weight is updated to improve performance. Finally, the ℓ_1 regularization is applied to the difference of intermediate activations for each batch, and the non-zero delta weights are updated to balance the performance and delta activation sparsity. For video inputs, the ℓ_1 regularization is applied to both delta task and temporal activations.

4.1. Single Image Evaluation For Multi-task

Model Baselines and Variants: We define the following baseline and model variants for the evaluation: (i) ‘**Multi-task learning (MTL)**’ represents a state-of-the-art (SOTA) baseline multi-task approach that comprises a shared encoder and multiple task-specific decoders which are jointly optimized. The current SOTA MTL baseline, InvPT [68], uses a transformer as the encoder, while others typically employ a CNN as the backbone. (ii) ‘**Single-task learning (ST)**’ has a common transformer backbone plus a task-specific small CNN head model structure for each task which is independently trained for task-specific parameters (for both backbone and head models) without using the delta weight and delta activation pruning methods. (iii) ‘**Ours + weight sharing**’ is similar to the ST model but adds delta weight pruning to share transformer weights between the base and sub-tasks. (iv) ‘**Ours + weight + act sharing**’ indicates our proposed method that also utilizes transformer activation sharing in the task domain.

Qualitative results: Figure 3 presents sample visualizations from the proposed model for the PASCAL-Context dataset. We compare these results with the current SOTA

InvPT [68] to demonstrate the superiority of our model. The visual comparison reveals that our model produces more accurate predictions compared to InvPT, particularly for semantic segmentation, human parsing, and edge detection tasks. Additional visualization results can be found in the supplemental material.

Quantitative results: Evaluation results for PASCAL-Context dataset and NYUD-v2 dataset are summarized in Table 1 (left) and Table 1 (right), respectively. The tables illustrate that our method outperforms prior approaches, including InvPT [68]. We compare the FLOPs, number of parameters, required memory to store model parameters, and performance of each task against the ‘ST’ model and the SOTA transformer-based MTL method InvPT in Table 2 (a) and (b). Notice that our base task (semantic segmentation) performance is similar to that of the ‘ST’ method since it does not use weight or activation sharing. For the PASCAL-Context dataset, our proposed model (‘weight + act sharing’) outperforms InvPT while reducing the FLOPs and parameters by **49.44%** and **74.0%**, respectively, compared to the ‘ST’ model. Table 2 (a) specifies the FLOPs required for each base and sub-task, indicating that adding a new sub-task requires only **37.6%** FLOPs of the ‘ST’ model. For the NYUD-v2 dataset, we first evaluate the model with a single image without enabling temporal activation reuse to isolate the gain of the task activation reuse strategy. Table 2 (b) shows that our method reduces the FLOPs and parameters by **40.5%** and **66.6%** compared to the ‘ST’ model while achieving comparable/better results than the InvPT baseline. To evaluate the required memory, we adopt the CSR (Compressed Sparse Row) method to store the sparse delta weight matrices. This approach efficiently stores only the non-zero elements and their corresponding positions, making it a suitable choice for our approach. The memory storage comparison results are presented in Table 2, further showcasing the memory-saving benefits of our approach. Figure 4 shows the overall delta weight and delta task activation sparsity for all sub-tasks in the Pascal-Context dataset (left) and the NYUD-v2 dataset (right). It reveals that human parsing and edge detection are more closely related to the base task, hence their sparsities are higher than other sub-tasks.

Ablation study: We explore the impact of ℓ_0 and ℓ_1 regularization coefficients on the performance, computation, and memory storage reduction for each task in both PASCAL-Context and NYUD-v2 datasets. To determine the optimal values of λ_w and λ_{a1} , we conduct a comprehensive hyperparameter search, sweeping λ_w in the range of $[1 \times 10^{-8}, 10 \times 10^{-8}]$ and λ_{a1} in the range of $[1 \times 10^{-10}, 10 \times 10^{-10}]$. Figure 5 shows the performance of the human parsing task with different values of λ_w and λ_{a1} . As expected, increasing λ_w saves more parameters but reduces accuracy by 0.30%. Similarly, increasing λ_{a1} saves

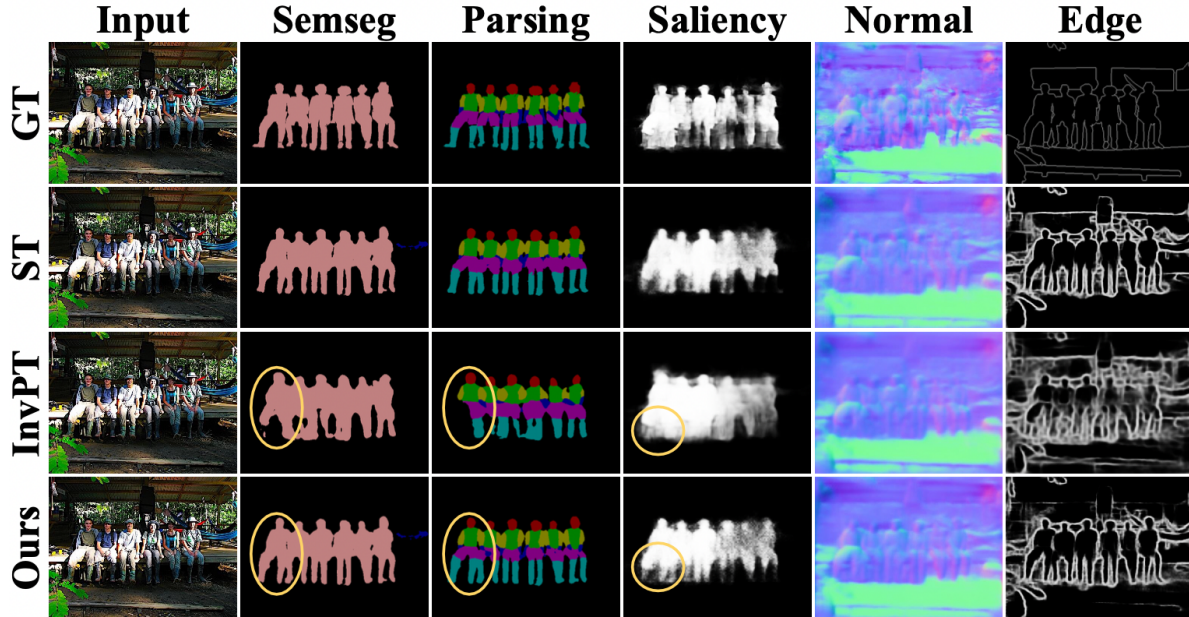


Figure 3: Qualitative comparison with the previous SOTA InvPT and ‘ST’ model on PASCAL-Context dataset. Examples of the regions where our model outperforms InvPT are shown with yellow circles.

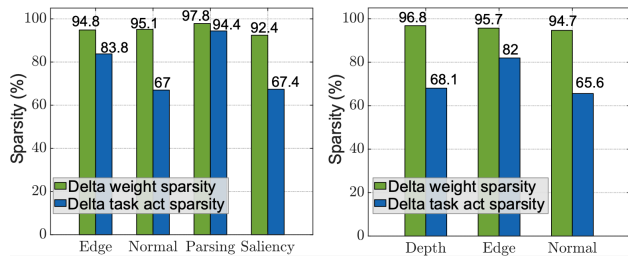


Figure 4: Overall delta weight and delta task activation sparsity for the Pascal-Context dataset (*left*) and the NYUD-v2 dataset (*right*).

more computations at the cost of decreased accuracy by 1.24%. For each sub-task, we select the optimal λ_w and λ_{a1} to balance computation/memory storage and task-specific performance.

In addition, we evaluate the impact of using a different backbone network by replacing the ViT-B transformer model with the Swin-B [37] model. This evaluation is performed for both the InvPT baseline and our proposed method with semantic segmentation (‘Semseg’) as the base task. Furthermore, we compare the performance of our method when human parsing (‘Parsing’) is selected as the base task, aiming to demonstrate the effectiveness of our approach when choosing an alternative task as the base. The results on the Pascal-context dataset are shown in Table 3. When all approaches use Swin-B, our method (weight & act sharing) with the ‘Semseg’ as the base task outperforms InvPT baseline, exhibiting significantly reduced FLOPs and parameters

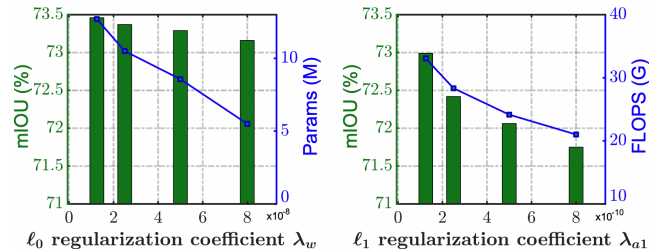


Figure 5: The impacts of λ_w and λ_{a1} on the human parsing task for Pascal-context dataset.

(by 45.73% and 71.07%) compared to the ‘ST’ model. Similarly, our method with the ‘Parsing’ as the base task surpasses the InvPT baseline, exhibiting substantial reductions in FLOPs and parameters (by 42.16% and 71.03%, respectively) compared to the ‘ST’ model. It is also evident that the ‘Semseg’ task delivers slightly better performance as the base task. We attribute this improvement to the fact that segmentation is a highly informative task, generating powerful activations and features that can be effectively generalized to other computer vision tasks.

4.2. Video Frame Evaluation

Now, we evaluate the temporal activation-sharing method using the NYUD-v2 dataset. As discussed in Sec 3.3, we first evaluate the sparsity ratios of all layers for a task and then determine the optimal layer boundary for sharing mode switching. Specifically, for task b_k , we leverage the task-

Table 3: Performance of our method when the base task is segmentation ‘Semseg’ or ‘Parsing’ on Pascal-context dataset using Swin-B as the transformer backbone (replacing ViT-B).

Method	Semseg mIoU ↑	Parsing mIoU ↑	Saliency maxF ↑	Normal mErr ↓	Boundary odsF ↑	Param (M)	FLOPs (G)
ST	79.1	71.6	84.6	13.5	76.0	516.1	444.3
InvPT	77.5	66.8	83.6	14.6	73.0	187.2	384.7
Ours w/Semseg base	79.1	70.0	84.2	14.3	74.5	149.3	241.1
Ours w/Parsing base	77.5	71.6	84.1	14.4	74.4	149.5	257.0

domain activation sharing for the first l^{bk} layers and subsequently switch to the temporal-domain activation sharing for the remaining layers. Figure 6 shows the sparsity comparison between the delta task and temporal activations for depth estimation, surface normal estimation, and edge detection tasks to determine the sharing mode switching boundary l^{bk} .

As the frames in the NYUD-v2 dataset are sparsely annotated (e.g., only every 20th frame is annotated), we evaluate the performance of annotated ground truth (GT) frames by considering all possible interval offsets between the start (keyframe) and GT frames within the range of $[0, 4]$ and record the averaged performance and FLOPs. To assess the impact of task and temporal activation combinations, we evaluate the following model variants: (i) ‘ST’, (ii) ‘InvPT’, (iii) ‘+ Task act’ using only task domain activation reuse for all frames; and (iv) ‘+ Task act + Temporal act’ which is the proposed model using the combination of temporal and task activation sharing strategy. The proposed model employs the same computation as the ‘+ Task act’ model for a keyframe (it appears up to 4 frames earlier than the GT frame) but reduces the computation of base and sub-tasks for non-keyframes. Table 4 compares the performance and FLOPs of these variants. We observe that the proposed approach significantly reduces FLOPs by **42.3%** and **65.7%** compared to ‘+ Task act’ and ‘ST’ models, respectively.

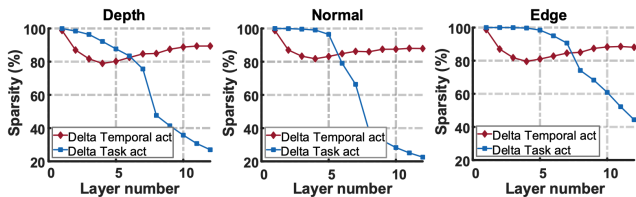


Figure 6: The sparsity comparison between delta task and temporal activation for NYUD-v2 dataset.

5. Discussion

Calculating the actual inference speed of deep learning models on available computing platforms is a critical factor in evaluating performance. However, due to the utilization of sparse matrix-matrix multiplications in our proposed approach, measuring the true speedup requires a specialized sparse linear algebra processor or hardware accelerator,

Table 4: Performance of our method using temporal and task activation re-use technique for NYUD-v2 dataset.

Method	Semseg mIoU ↑	Depth RMSE ↓	Normal mErr ↓	Boundary odsF ↑	Param (M)	FLOPs (G)
ST	50.48	521.5	17.26	77.84	366.64	314.14
InvPT	50.30	536.7	19.00	77.6	160.57	229.43
Ours + Task act	50.46	533.2	18.42	77.89	122.46	186.88
Ours + Task act + Temporal act	50.40	532.41	18.42	77.79	122.46	107.87

which is not yet readily available for commercial platforms. As a practical alternative, we analyze the required FLOPs to estimate the speedup potential. While FLOPs provide a reasonable comparison metric, we acknowledge that they may not fully represent the actual speed gains achieved. This calls for future work to develop a sparsity-aware transformer accelerator that will allow us to quantify the true speedup achieved by our approach.

6. Conclusion

This paper presents a novel computation- and parameter-sharing scheme for transformer-based multiple visual tasks that are concurrently performed on the same input. Motivated by recent transfer learning techniques, our scheme reuses the weights and activations of the base task by training sub-tasks with sparse weight and activation differences via ℓ_0 and ℓ_1 regularization. As a result, the activations from the base task can be shared with all sub-tasks, reducing both parameter and computation redundancy significantly. Additionally, the proposed scheme is extended to video inputs to further reduce computation redundancy in the temporal domain. Evaluation results confirm that our method attains better/comparable performance with fewer parameters and FLOPs than state-of-the-art multi-task learning methods.

Acknowledgements. This work was supported in part by COGNISENSE, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. The authors would like to thank Morteza Tavakoli Taba, Bowen Liu, and Changwoo Lee at the University of Michigan for their valuable discussions.

References

- [1] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3286–3295, 2019. 3
- [2] Deblina Bhattacharjee, Tong Zhang, Sabine Süstrunk, 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, 3
- [3] David Brüggenmann, 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. 2, 6
- [4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3
- [5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 3
- [6] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 1, 2
- [7] Hanqing Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 1, 3
- [8] 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
- [9] 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. 2
- [10] Ziqian Chen, Ling-Yu Duan, Shiqi Wang, Yihang Lou, Tiejun Huang, Dapeng Oliver Wu, and Wen Gao. Toward knowledge as a service over networks: A deep learning model communication paradigm. *IEEE Journal on Selected Areas in Communications*, 37(6):1349–1363, 2019. 2
- [11] Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretschmar, 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. 2
- [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7
- [13] 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. In *International Conference on Learning Representations*, 2021. 1, 3, 4, 7
- [14] Simon Shaolei Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2020. 3
- [15] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015. 2
- [16] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi-task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117, 2004. 1, 2
- [17] Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516, 2021. 2
- [18] 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/CVF conference on computer vision and pattern recognition*, pages 3205–3214, 2019. 1, 2
- [19] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2
- [20] Mitchell A Gordon, Kevin Duh, and Nicholas Andrews. Compressing bert: Studying the effects of weight pruning on transfer learning. *ACL 2020*, page 143, 2020. 3
- [21] Demi Guo, Alexander Rush, and Yoon Kim. Parameter-efficient transfer learning with diff pruning. In *Annual Meeting of the Association for Computational Linguistics*, 2021. 2, 3, 5
- [22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3
- [23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2
- [24] Xuehai He, Chunyuan Li, Pengchuan Zhang, Jianwei Yang, and Xin Eric Wang. Parameter-efficient fine-tuning for vision transformers. *arXiv preprint arXiv:2203.16329*, 2022. 3
- [25] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4
- [26] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 2, 3
- [27] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 3
- [28] Zhiqi Huang, Lu Hou, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. Ghostbert: Generate more features with cheap operations for bert. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6512–6523, 2021. 3
- [29] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter

- layers. *Advances in Neural Information Processing Systems*, 34:1022–1035, 2021. 3
- [30] 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 conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 2
- [31] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019. 3
- [32] Abhishek Kumar and Hal Daumé III. Learning task grouping and overlap in multi-task learning. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, pages 1723–1730, 2012. 2
- [33] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pre-trained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. 3
- [34] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Universal representations: A unified look at multiple task and domain learning. *arXiv preprint arXiv:2204.02744*, 2022. 1
- [35] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019. 2
- [36] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*, 2019. 3
- [37] 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, 4, 8
- [38] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 1
- [39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *iclr*, 2019, 2017. 7
- [40] Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through L₀ regularization. In *International Conference on Learning Representations*, 2018. 5
- [41] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017. 5
- [42] 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, 2, 6
- [43] 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. 1, 2, 6
- [44] Eslam Mohamed and Ahmad El Sallab. Spatio-temporal multi-task learning transformer for joint moving object detection and segmentation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 1470–1475. IEEE, 2021. 1, 3
- [45] Hyeonwoo Noh, Taehoon Kim, Jonghwan Mun, and Bohyung Han. Transfer learning via unsupervised task discovery for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8385–8394, 2019. 3
- [46] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pages 4055–4064. PMLR, 2018. 3
- [47] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021. 3
- [48] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in neural information processing systems*, 32, 2019. 3
- [49] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 1
- [50] Victor Sanh, Thomas Wolf, and Alexander Rush. Movement pruning: Adaptive sparsity by fine-tuning. *Advances in Neural Information Processing Systems*, 33:20378–20389, 2020. 3
- [51] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Video multi-task transformer network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1, 3
- [52] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 6
- [53] Asa Cooper Stickland and Iain Murray. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *International Conference on Machine Learning*, pages 5986–5995. PMLR, 2019. 3
- [54] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2019. 3
- [55] Bethan Thomas, Samuel Kessler, and Salah Karout. Efficient adapter transfer of self-supervised speech models for automatic speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7102–7106. IEEE, 2022. 3
- [56] Katrin Tomanek, Vicky Zayats, Dirk Padfield, Kara Vaillancourt, and Fadi Biadsy. Residual adapters for parameter-efficient asr adaptation to atypical and accented speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6751–6760, 2021. 3
- [57] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training

- data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 3
- [58] Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862, 2020. 3
- [59] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *European Conference on Computer Vision*, pages 527–543. Springer, 2020. 1, 2, 6
- [60] 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*, 2021. 2
- [61] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2020. 1
- [62] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 3
- [63] Ziheng Wang, Jeremy Wohlwend, and Tao Lei. Structured pruning of large language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6151–6162, 2020. 5
- [64] Sen Wu, Hongyang R Zhang, and Christopher Ré. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations*, 2020. 3
- [65] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Padnet: 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, 2, 6
- [66] Yanhua Yang, Cheng Deng, Shangqian Gao, Wei Liu, Dapeng Tao, and Xinbo Gao. Discriminative multi-instance multitask learning for 3d action recognition. *IEEE Transactions on Multimedia*, 19(3):519–529, 2016. 2
- [67] Yanhua Yang, Cheng Deng, Dapeng Tao, Shaoting Zhang, Wei Liu, and Xinbo Gao. Latent max-margin multitask learning with skelets for 3-d action recognition. *IEEE transactions on cybernetics*, 47(2):439–448, 2016. 2
- [68] Hanrong Ye and Dan Xu. Inverted pyramid multi-task transformer for dense scene understanding. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 514–530. Springer, 2022. 1, 3, 6, 7
- [69] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, 2022. 3
- [70] Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11197–11206, 2020. 1
- [71] Matthew D Zeiler. Adadelata: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012. 7
- [72] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 3
- [73] Xiaoya Zhang, Ling Zhou, Yong Li, Zhen Cui, Jin Xie, and Jian Yang. Transfer vision patterns for multi-task pixel learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 97–106, 2021. 1
- [74] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021. 1
- [75] 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/CVF conference on computer vision and pattern recognition*, pages 4106–4115, 2019. 1, 2, 6
- [76] Mengjie Zhao, Tao Lin, Fei Mi, Martin Jaggi, and Hinrich Schütze. Masking as an efficient alternative to finetuning for pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2226–2241, 2020. 3
- [77] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 4514–4523, 2020. 1, 2, 6
- [78] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 1