Supplementary Information

This document provides more details about our proposed method and comparison.

1.1. Detailed Training Settings

We list detailed training information in Table 1, including data augmentation, hyperparameters, and training devices.

NYUD-v2 [1] contains various indoor scenes such as offices and living rooms with 795 training and 654 testing images. It provides different dense labels, including semantic segmentation, monocular depth estimation, surface normal estimation, and object boundary detection. **PASCAL-Context** is formed from the PASCAL dataset [2]. It has 4,998 images in the training split and 5,105 in the testing split, covering both indoor and outdoor scenes. This dataset provides pixel-wise labels for semantic segmentation, human parsing, and object boundary detection. Additionally, [3] generates surface normal and saliency labels for this dataset.

Table 1. Training Hyperparameters.

Training set	NYUD-v2	PASCAL-Context
# images	795 training	4,998 training
Image size	Around 448×576	Around 512×512
Data augment.	Crop, h-flip, Scale	Crop, h-flip, Scale
Optimizer	Adam	Adam
Learning rate	2×10^{-5}	2×10^{-5}
LR schedule	Polynomial	Polynomial
Batch size	2	2
# iterations	160k	160k

1.2. Detailed Neural Network Architectures

The EVA-02 pre-trained vision backbone used in our framework is a four-stage plain ViT with modified blocks. The detailed architecture of its Transform Vision (TrV) block is depicted in Fig. 1. The block incorporates a 2D rotary position embedding (RoPE) [4] for injecting positional information, a gated linear unit (GLU) [5, 6] with a sigmoid linear unit (SiLU) [7] as the feedforward network, and sub-layer normalization (sub-LN) [?] as the normalization layer. The pre-training strategies for EVA-02 are detailed in [8]. We utilize the pre-trained weights directly available via timm[1].

1.3. Quantitative Results

We present the BD-Rate, BD-Accuracy, and the number of trainable parameters for various methods in Tab. 2 and Tab. 3. Note that when calculating BD-Rate and BD-Accuracy, only the overlapping regions are considered. Our proposed method exhibits substantial improvements over both the anchor method and the Pre-processor method in terms of BD-Rate and BD-Accuracy across all tasks, while maintaining comparable performance. For instance, the proposed DoRA FT achieves an impressive -80.03% BD-Rate and 17.91% BD-mIoU compared to the VVC + Full FT Baseline anchor. In contrast, the Pre-processor + DoRA FT Baseline achieves only -39.60% BD-Rate and 6.27% BD-mIoU on the NYUD-v2 semantic segmentation task. Furthermore, the total trainable parameters for these methods are 84M and 87M, respectively, underscoring the superior parameter efficiency of the proposed framework.

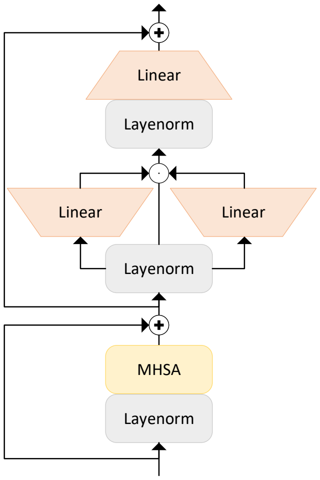


Fig. 1. Transform Vision (TrV) block

1.4. Task Interaction Results

In the main experiment, separate decoders are used for different tasks. However, recent work [9] has demonstrated that modeling task relationships can significantly enhance performance. To explore this, we propose a straightforward approach to integrate task interactions. Specifically, task features are concatenated along the spatial dimension in the decoder to form a multitask feature representation, $\mathbf{F}_s \in \mathbb{R}^{T\times H_s \times W_s \times C_s}$, where T, H_s, W_s, and C_s represent the number of tasks, spatial dimensions, and channels at stage s, respectively. Using this representation, the query, key, and value projections are defined as follows: $\mathbf{Q}_s = \mathbf{W}_s^q(\text{Conv}(\mathbf{F}_s)) \in \mathbb{R}^{\frac{TH_sW_s}{4}\times C_s}$, $\mathbf{K}_s = \mathbf{W}_s^k(\text{Conv}(\mathbf{F}_s, k_s))$, $\mathbf{V}_s = \mathbf{W}_s^v(\text{Conv}(\mathbf{F}_s, k_s))$, where $\mathbf{K}_s, \mathbf{V}_s \in \mathbb{R}^{\frac{TH_sW_s}{k_s^2}\times C_s}$. The self-attention score matrix is then computed as: $\mathbf{A}_s = \mathbf{Q}_s\mathbf{K}_s^T \in \mathbb{R}^{\frac{TH_sW_s}{4}\times\frac{TH_sW_s}{k_s^2}}$. In this experiment, the encoder from the main experiment is reused, and only the decoder is retrained.

All other configurations remain consistent with the main method. Additionally, our low-complexity decoder facilitates efficient self-attention across both the spatial and task dimen-

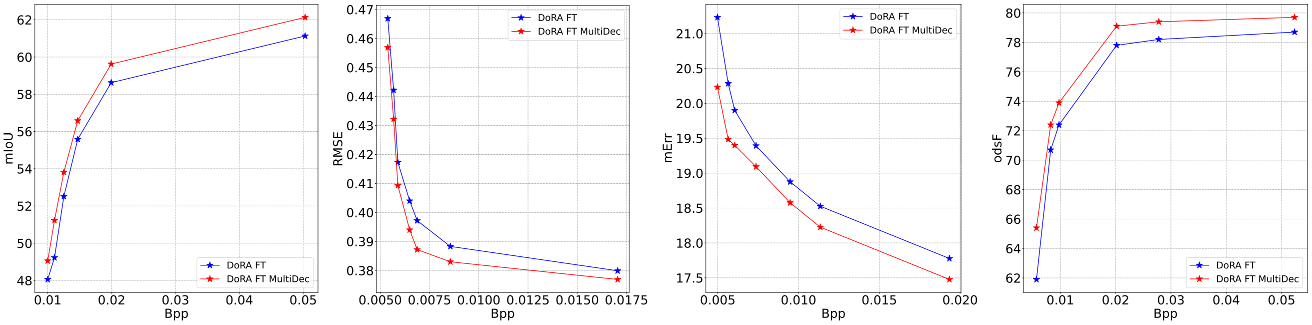
[1] https://huggingface.co/timm/eva02_large_patch14_448.mim_m38m_ft_in22k

Table 2. Comparison of rate-accuracy performance and the number of trainable parameters under different machine vision tasks and different methods on the NYUD-v2 dataset. VVC + Full FT Baseline is used as the anchor to calculate BD-Rate and BD-Accuracy. Arrows indicate whether lower is better (↓) or higher is better (↑).

Method	**Semseg**		**Depth**		**Normal**		**Boundary**		**Total Trainable**	**Backbone Trainable**
	BD-Rate ↓	**BD-mIoU ↑**	**BD-Rate ↓**	**BD-RMSE ↓**	**BD-Rate ↓**	**BD-mErr ↓**	**BD-Rate ↓**	**BD-odsF ↑**	**Params ↓ (M)**	**Params ↓ (M)**
VVC + Full FT Baseline	0%	0%	0%	0	0%	0	0%	0%	330M	305M
VVC + DoRA FT Baseline	-27.16%	4.20%	-37.66%	-0.06	7.92%	0.34	33.64%	-3.31%	27M	1M
Pre-processor + Full FT Baseline	-18.90%	2.65%	-19.61%	-0.03	-23.85%	-1.20	-17.70%	1.90%	390M	305M
Pre-processor + DoRA FT Baseline	-39.60%	6.27%	-50.87%	-0.09	-17.57%	-0.84	10.78%	-1.13%	87M	1M
Full FT	-76.68%	19.79%	-87.66%	-0.92	-88.14%	-8.68	-77.74%	13.46%	387M	305M
DoRA FT	-80.03%	17.91%	-97.08%	-0.71	-89.39%	-8.97	-67.27%	8.84%	84M	1M

Table 3. Comparison of rate-accuracy performance and the number of trainable parameters under different machine vision tasks and different methods on the PASCAL- Context dataset. VVC + Full FT Baseline is used as the anchor to calculate BD-Rate and BD-Accuracy. Arrows indicate whether lower is better (↓) or higher is better (↑).

Method	**Semseg**		**Parsing**		**Saliency**		**Normal**		**Boundary**	
	BD-Rate ↓	**BD-mIoU ↑**	**BD-Rate ↓**	**BD-mIoU ↑**	**BD-Rate ↓**	**BD-maxF ↑**	**BD-Rate ↓**	**BD-mErr ↓**	**BD-Rate ↓**	**BD-odsF ↑**
VVC + Full FT Baseline	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
VVC + DoRA FT Baseline	-18.61%	3.89%	-15.42%	3.13%	13.71%	-0.98%	-19.04%	-0.67%	-1.83%	0.25%
Pre-processor + Full FT Baseline	-20.05%	3.66%	-21.45%	4.25%	-32.47%	2.81%	-37.47%	-1.54%	-22.04%	3.79%
Pre-processor + DoRA FT Baseline	-33.85%	6.89%	-34.78%	6.80%	-22.03%	1.85%	-52.00%	-2.17%	-21.27%	3.46%
Full FT	-85.57%	29.75%	-91.16%	26.94%	-92.30%	66.44%	-92.96%	-8.55%	-86.20%	23.13%
DoRA FT	-85.59%	26.04%	-92.19%	23.64%	-92.07%	32.58%	-93.24%	-8.82%	-82.46%	24.01%



(a) Semseg

(b) Depth

(c) Normal

(d) Boundary

Fig. 2. Comparison of separate coder and multitask decoder on NYUD-v2 dataset.

sions, maintaining acceptable computational complexity. The results, shown in Fig. 2, clearly demonstrate that incorporating task interactions yields significant performance improvements. These findings highlight the potential of task interaction modeling and underscore its importance as a promising area for further research.

2. REFERENCES

[1] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor segmentation and support inference from rgbd images," *Proceedings of the European Conference on Computer Vision*, pp. 746–760, 2012.

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010.

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

[4] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, pp. 127063, 2024.

[5] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier, "Language modeling with gated convolutional networks," *Proceedings of the International Conference on Machine Learning*, pp. 933–941, 2017.

[6] Noam Shazeer, "Glu variants improve transformer," 2020.

[7] Dan Hendrycks and Kevin Gimpel, "Gaussian error linear units (gelus)," 2016.

[8] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao, "Eva-02: A visual representation for neon genesis," *Image and Vision Computing*, vol. 149, pp. 105171, 2024.

[9] Anonymous, "Which tasks should be compressed together? a causal discovery approach for efficient multi-task representation compression," 2024.