Region-aware Distribution Contrast: A Novel Approach to Multi-Task Partially Supervised Learning

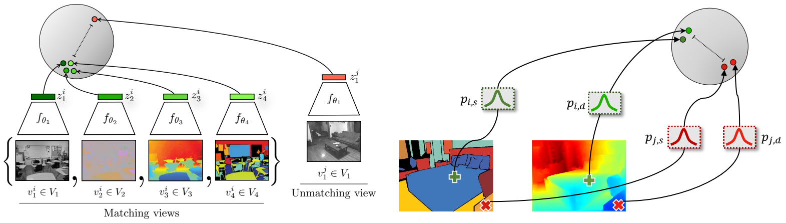
Meixuan Li, Tianyu Li, Guoqing Wang*,
Peng Wang, Yang Yang, and Jie Zou

University of Electronic Science and Technology of China (UESTC)
limeixuan0801@126.com, {cosmos.yu, p.wang6}@hotmail.com,
{gqwang0420, yang.yang, jie.zou}@uestc.edu.cn

Abstract. In this study, we address the intricate challenge of multi-task dense prediction, encompassing tasks such as semantic segmentation, depth estimation, and surface normal estimation, particularly when dealing with partially annotated data (MTPSL). The complexity arises from the absence of complete task labels for each training image. Given the inter-related nature of these pixel-wise dense tasks, our focus is on mining and capturing cross-task relationships. Existing solutions typically rely on learning global image representations for global cross-task image matching, imposing constraints that, unfortunately, sacrifice the finer structures within the images. Attempting local matching as a remedy faces hurdles due to the lack of precise region supervision, making local alignment a challenging endeavor. The introduction of Segment Anything Model (SAM) sheds light on addressing local alignment challenges by providing free and high-quality solutions for region detection. Leveraging SAM-detected regions, the subsequent challenge lies in aligning the representations within these regions. Diverging from conventional methods that directly learn a monolithic image representation, our proposal involves modeling region-wise representations using Gaussian Distributions. Aligning these distributions between corresponding regions from different tasks imparts higher flexibility and capacity to capture intra-region structures, accommodating a broader range of tasks. This innovative approach significantly enhances our ability to effectively capture cross-task relationships, resulting in improved overall performance in partially supervised multi-task dense prediction scenarios. Extensive experiments conducted on two widely used benchmarks underscore the superior effectiveness of our proposed method, showcasing state-of-the-art performance even when compared to fully supervised methods.
https://github.com/HereNowL/Region-aware-Distribution-Contrast

Keywords: Multi-task Learning · Partially Supervised Learning · Scene Understanding · Contrastive Learning

* Corresponding author.



(a) Contrastive Multiview Coding

(b) Region-wise Distribution Contrast

Fig. 1: Multiview consistency and region consistency. (a) Illustration of contrastive multiview consistency [38]. (b) Illustration of region consistency, where $p_{i,s}$ and $p_{j,s}$ represent the region distribution of semantic segmentation, $p_{i,d}$ and $p_{j,d}$ denotes the region distribution of depth estimation.

1 Introduction

With the rapid advancement of deep learning, dense prediction tasks, including semantic segmentation, depth estimation, and surface normal estimation, have witnessed remarkable progress [11, 16, 30, 34, 47]. Considering the inherent interdependence among these dense prediction tasks [45,46], there has been a growing interest in employing unified multi-task learning networks to jointly tackle different dense prediction tasks [7,22,28,29,31,32,39,43,50]. In contrast to inefficient single-task learning networks, multi-task learning networks effectively acquire shared features, enabling them to capture more generalizable information across various tasks while avoiding redundant training [39,43,50].

Recent advancements in multi-task dense prediction methods have primarily focused on two crucial aspects. The first aspect is dedicated to designing intricate network architectures to achieve effective multi-task learning [29, 32, 39, 43, 50], while the second aspect is devoted to devising optimal balancing strategies for loss functions in multi-task learning [7, 14, 22, 28, 29], aiming to mitigate the negative transfer across tasks. These directions mainly focus on fully supervised multi-task learning, where complete task labels are available for all training samples. However, acquiring pixel-level labels for dense prediction tasks is resource-intensive, particularly with multiple visual tasks. It is common to encounter scenarios where labels for certain tasks are missing or unreliable in some training samples [25], leading to a new research direction known as multi-task partially supervised learning (MTPSL).

Since not all task labels are available for each training image in MTPSL, it is natural to formulate a multi-task learning framework which can leverage the implicit relationship between tasks with and without explicit labels. Dense prediction tasks, such as semantic segmentation, depth estimation, and normal prediction, exhibit a complementary co-existence relationship [37,45,46], where each task can serve as auxiliary information for others, offering valuable insights reciprocally. While existing approaches map predictions into a joint vector-wise semantic space for task alignment [25], we argue that this overlooks the underlying structural information inherent in the original image. A global vector might be insufficient to characterize the information in the entire scene, limiting its effectiveness in constraining cross-task relationships. Local matching, as shown in Fig. 1, with its intrinsic local properties, serves as a potential remedy. Nonetheless, the absence of precise region supervision presents a fundamental obstacle, thereby intensifying the challenges associated with its implementation.

To tackle the outlined challenges, we introduce region-based cross-task alignment by leveraging the robust region segmentation capabilities inherent in Segment Anything Model (SAM) models. Utilizing well-segmented image regions facilitates the abstraction of region-level information, enabling the addition of cross-task consistency constraints. However, a subsequent crucial challenge arises in effectively representing these regional features to ensure the consistency and accuracy of cross-task alignment. In contrast to conventional methods that directly learn a unified image representation, we employ Gaussian distributions to model region representations, enhancing flexibility and improving the ability to capture intra-region structures. By adapting Gaussian distributions between corresponding regions from different tasks, our approach can be more broadly applicable to diverse task types, such as pixel-level classification tasks (e.g., semantic segmentation) and pixel-level regression tasks (e.g., depth estimation). Subsequently, we introduce distributional contrastive learning to explicitly perform region-wise cross-task alignment. This approach serves as a robust constraint, offering a meaningful way to mine relationships between tasks and enhance dense prediction results, particularly for those tasks lacking pixel-wise labels.

In summary, our work masks the following contributions:

- We tackle the challenge of multi-task partially supervised (MTPSL) dense prediction from a fresh perspective. Our approach involves extracting cross-task local alignment through the utilization of SAM's easily obtainable local regions. This innovative strategy has demonstrated notable in alleviating label shortages.
- We propose a novel region distribution contrast method for local alignment, which offers increased flexibility, robustness, and wide applicability across multiple tasks.
- We validate the effectiveness of our proposed method through extensive experiments conducted on two widely used benchmark datasets. Notably, our approach achieves state-of-the-art performance in partially supervised learning, while also showing great potential in fully supervised learning.

2 Related Work

Multi-task Supervised Learning. Multi-task learning utilizes a single model to simultaneously handle multiple visual tasks, offering advantages such as faster inference speed and more efficient utilization of input data. Some approaches [13, 32, 48] focus on designing the encoder of multi-task network, while other approaches pay more attention to designing complex decoders to generate task-specific predictions by utilizing shared features [2, 39, 43, 50]. Uncertainty [22] obtain task loss weights by considering the heteroscedastic uncertainty of each task, Grad-Norm [7] directly adjusts gradients to balance task losses, and DWA [14] learns time-varying average task weights by considering the loss variation rate of each task. It is important to note that these methods operate under a significant and strong assumption, namely that they are developed and studied based on the availability of labels for all tasks across all images.

Multi-task Semi-supervised Learning and Partially-supervised Learning. Learning multi-task models on fully annotated data requires a large-scale labeled dataset, and the cost of collecting sufficient labeled data can be high. Therefore, two common scenarios are worthy of note in multi-task learning. One is semi-supervised multi-task learning, where the dataset consists of limited annotation information for all tasks and a large amount of unlabeled data. A bunch of methods have been proposed for semi-supervised multi-task learning [8,18,19,24,27]. In recent works [8,18,19,24], regularization terms are applied to unlabeled samples from each task to encourage consistent predictions when the input is perturbed. Another scenario involves partial-supervision multi-task learning, where the dataset lacks sufficient labels for each task, and not all tasks have labels for every image. Li *et al.* [25] maps tasks into high-dimensional vectors for cross-task alignment. However, vector representations are insufficient and do not capture local-level alignment. In this paper, we propose leveraging distributions as a substitute for vector representations of task features and employing contrastive learning to achieve region-level alignment across tasks.

Contrastive Learning for Vision. Recently, contrastive learning has made significant advancements in unsupervised learning [5,6,15,33,42]. DenseCL [40] and RegionContrast [17] have been utilized for semantic segmentation tasks using contrastive learning at the pixel and region levels respectively, achieving excellent results. For depth estimation, WCL [12] transformed depth from continuous values to discrete values and constructed contrastive losses based on windows to form positive and negative sample pairs. CMC [38] was the first to propose that different viewpoints of the same image should be mapped to nearby positions in a high-dimensional space, while different images should be mapped to distant positions, thus ensuring multiview consistency. Inspired by CMC, we believe that multiple viewpoints within the same local region should also exhibit consistency, while differing from other regions. In this paper, we leverage this idea to achieve region-level contrast for MTPSL.

Gaussian Distribution. Recently, Gaussian distribution has been widely employed in computer vision to characterize the distribution of features [3,4, 20,26,41,49]. GMMSeg [26] utilized the Expectation-Maximization algorithm to construct Gaussian Mixture Models for each class in semantic segmentation, capturing class-conditional densities. AGMM [41] constructed an adaptive Gaussian Mixture Model for sparse annotation in semantic segmentation by incorporating labeled pixels and their similar unlabeled counterparts. DUL [3] employed mean and variance to characterize the distribution of faces for face recognition. Gaussian distributions have the capability to probabilistically describe the distribution characteristics of different visual tasks. However, to the best of our knowledge, Gaussian distributions have not been applied in MTPSL.

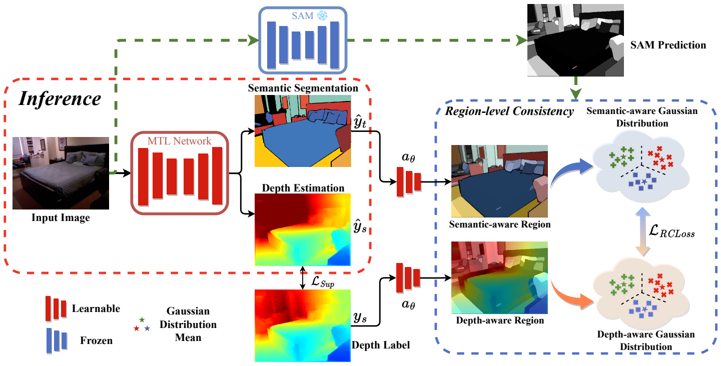


Fig. 2: Illustration of region-aware distribution contrast learning method for MTPSL. During training, supervised constraints L_{Sup} are applied to the annotated task s. For task s and unlabelled task t, the a_θ map the true label y_s and the prediction $\hat{y}_t$ in the high-dimensional space respectively and then model the region of task-specific features extracted using SAM as a Gaussian distribution. Contrastive learning is then employed to minimize the distance between distributions of the same region across different tasks while maximizing the distance to distributions of other regions.

3 Method

3.1 Preliminaries

Problem Setup. Considering the problem of multi-task learning involving K tasks, where $K \geq 2$. A training dataset $S = \{x_i\}_{i=1}^{N}$ is given with N partially labelled samples, indicating that for each training sample x_i, only a subset of the K tasks are provided true labels. Let P_i represents the number of labeled tasks for x_i and Q_i represents the number of unlabeled tasks, such that $P_i + Q_i = K$. When P_i equals K for all $x_i \in S$, it indicates fully supervised multi-task learning. Conversely, when Q_i equals K for all $x_i \in S$, it implies that no task labels are available, leading to unsupervised multi-task learning. In this paper, we focus on addressing partially supervised multi-task learning, where each training image x_i can obtain labels for at least one task ($P_i \geq 1$).

3.2 Overview of Our Method

As shown in Fig. 2, our method focuses on utilizing Gaussian distributions to represent features in local regions for contrastive learning, achieving better distribution consistency across tasks at the region-level for MTPSL. Assuming that task s has label while task t does not, our objective is to achieve alignment between the true label y_s of task s and the prediction $\hat{y}_t$ of task t at the region level using contrastive learning.

In the training stage, to accurately and efficiently obtain region-level features for prediction, we simultaneously feed the input image to both the multi-task learning (MTL) network and the pre-trained SAM. Predictions are obtained

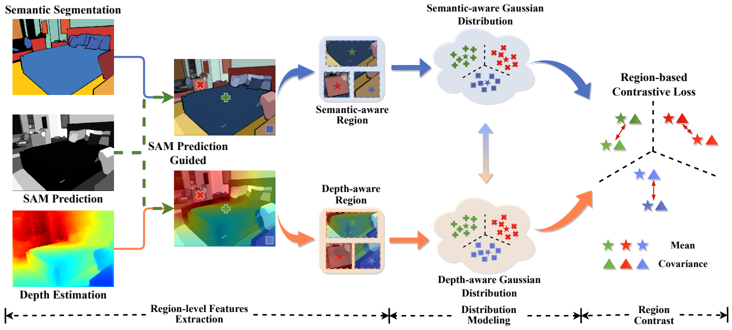


Fig. 3: Illustration of region-level cross-task consistency. Initially, SAM predictions are employed to extract regions from the features of the particular task. Following that, these regions are modeled as Gaussian distributions. Finally, contrastive learning is utilized to minimize the distance between regions of the same region across different tasks and maximize the distance to regions of other regions.

from both MTL network and SAM. For tasks with labeled data, we apply fully-supervised loss L_{Sup} as explicit training supervision. For tasks without label, an auxiliary network a_θ maps the unlabelled task prediction $\hat{y}_t$ and the annotated task label y_s to the same high-dimensional feature map space, and use L_{RC} to achieve cross-task consistency at the region level. Specifically, We first extract region-wise features in $\hat{y}_t$ and y_s based on the segmented regions identified by SAM, then model them as Gaussian distributions. After acquiring the Gaussian distributions for features in each region of each task, we utilize contrastive learning to minimize the distance between the Gaussian distributions of features in the same region across different tasks and maximize the distance from the Gaussian distributions of features from different regions. Finally, the optimization of our method can be defined as:

$$Loss = L_{Sup} + L_{RC}. \tag{1}$$

In the inference stage, as illustrated in Fig. 2, the input image is fed only into the MTL network to generate predictions for multiple tasks, notably, with SAM no longer being utilized.

3.3 Region-aware Contrastive Learning

In this section, we explain the implementation details to achieve region-level cross-task consistency.

Region-level Features Extraction. To realize the region-level cross-task consistency, it is essential to informatively choose regions for feature extraction. Considering the robust capabilities of SAM for image segmentation [23], we employ it to generate the regions for each training sample, leveraging its ability to provide valuable prior knowledge. As shown in Fig. 2, to obtain fine-grained semantics and fine-grained edges of regions, we fed the image into MTL network and pre-trained SAM simultaneously. SAM performs fine-grained segmentation of each object in the image, and its output masks contain the ID of each area and its corresponding position in the image. Inspired by [21], for convenient region-wise feature extraction in task-specific feature maps, we compute the predicted masks of SAM as a grayscale image, where each pixel value corresponds to the area ID returned by SAM, indicating its category. Leveraging the grayscale masks of SAM, we split the image into different regions and extract region-wise features within feature maps of different tasks.

Gaussian Distribution Modeling. After obtaining the region-wise features for each task, we consider modeling the region-wise features as Gaussian distributions. The utilization of Gaussian distributions represents the features of a region in a probabilistic manner, which provides a more comprehensive depiction of the region's variations and compensates for the shortcomings of pixel-level hard alignment. Specifically, we model each region r_i from the task-specific feature map f_t as a Gaussian distribution:

$$p_{i,t} = p(r_i|f_t) = N(r_i; \boldsymbol{\mu}_{i,t}, \boldsymbol{\Sigma}_{i,t}), \tag{2}$$

where $\boldsymbol{\mu}_{i,t}$ and $\boldsymbol{\Sigma}_{i,t}$ represent mean and covariance matrix respectively, calculated from the region r_i in the feature map f_t, as shown in Fig. 3.

Region Distribution Contrast. Once we have obtained the Gaussian distributions for each region, our objective is to utilize contrastive learning to minimize the distance between Gaussian distributions of the same region across different tasks and maximize the distance between Gaussian distributions of different regions. Hence, the region-wise Gaussian distribution contrastive loss for the region distribution $p_{i,t}$ can be defined as follows:

$$L_{i,t}^{NCE} = -\textstyle\sum_{k_+} \log \frac{sim\left(p_{i,t}, k_+\right)}{sim\left(p_{i,t}, k_+\right) + \sum_{k_-} sim\left(p_{i,t}, k_-\right)}, \tag{3}$$

$$sim(p_{i,t}, k) = exp(-W_{distance}(p_{i,t}, k)/\tau), \tag{4}$$

where $k_+ \in \{p_{i,s}\}_{s \neq t}$, $k_- \in \{p_{j,*}\}_{j \neq i}$, $*$ denotes any task, sim denotes the exponential equation of the Wasserstein distance, and τ is the temperature parameter. Wasserstein distance [35] is a metric used to quantify the dissimilarity between two probability distributions and well-suited for accurately measuring the distance between discrete and continuous distributions, as it accounts for the underlying structure and geometry of the data. It measures the minimum amount of work required to transform one distribution into another and can be calculated as follows:

$$\begin{aligned} W_{distance}(p_{i,t}, k) = {} & \left\|\boldsymbol{\mu}_{i,t} - \boldsymbol{\mu}_k\right\|^2 + \operatorname{Tr}\left(\boldsymbol{\Sigma}_{i,t}\right) \\ & + \operatorname{Tr}\left(\boldsymbol{\Sigma}_k\right) - 2\operatorname{Tr}\left(\left(\boldsymbol{\Sigma}_{i,t}\boldsymbol{\Sigma}_k\right)^{1/2}\right), \end{aligned} \tag{5}$$

where Tr represents the trace of the matrix, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ represent the mean and covariance matrix of sample k. Based on L^{NCE}, the region-wise contrastive loss

L_{RC} can be defined as follows:

$$L_{RC} = \frac{1}{KM} \sum_{t}^{K} \sum_{i}^{M} L_{i,t}^{NCE}, \tag{6}$$

where M is the number of regions in the feature map t.

By minimizing L_{RC}, the network learns to contrast the distributions from the same region with those from different regions, thereby enabling better cross-task consistency at the region level.

Alternative Local Extraction Strategy. Here we explore an alternative local extraction strategy to investigate local cross-task consistency. To extract local information from the feature map, the simplest approach is to divide the map into patches, where patches at the same position serve as positive samples, and their distances are minimized. Conversely, patches from other positions act as negative samples, and their distances are maximized. We refer to this local extraction strategy as patch-aware contrast, which will be compared to the region-aware contrast method in Sec. 4.

Alternative Local Contrast Strategy. Alternatively, the method of modeling regions as Gaussian distributions and performing contrastive learning can be replaced with other contrastive strategies. One strategy is pixel-to-pixel level region contrast, where each pixel within the region is pulled closer to the corresponding pixel in the positive sample region and pushed far apart from each pixel in the negative sample region. We refer to this as pixel contrast. Another strategy is to map the region to a vector and bring it closer to the vector of the positive sample region while pushing it away from the vector of the negative sample region. This strategy is referred to as vector contrast. We compare these alternative methods in Sec. 4.

4 Experiments

Datasets. We evaluated our method on two standard dense prediction multi-task datasets: NYU-V2 [36] and Cityscapes [9]. NYU-V2 is an indoor dataset of natural scenes, including three dense prediction tasks: semantic segmentation, depth estimation, and surface normal estimation. The semantic segmentation task in NYU-V2 contains 13 categories, depth estimation is provided by Microsoft Kinect, and surface normal estimation is provided by [10]. As [25,29], all images were resized to 288x384 for training and evaluation purposes. Cityscapes is a well-known dataset for autonomous driving, consisting of outdoor street scenes with two tasks: semantic segmentation and depth estimation. Following [25, 29], we used 7-class semantic segmentation annotations for evaluation and resized all images to 128x256 to improve training speed.

Experimental Setting. Following [25], we evaluated our method in two partially supervised multi-task settings: the one label setting and the random label setting. In the one label setting, each image is assigned a random task with a label. In the random label setting, each image is assigned random labels for

Table 1: Quantitative comparison on the NYU-V2 dataset. ↑ (↓) denotes that, larger (smaller) values lead to better quality. * represents the available partially supervised SOTA. * represents the available fully supervised SOTA. + denotes the percentage improvement over the SOTA performance, corresponding to ↑ (↓). The bold denotes the best.

Label setting	Method	Backbone	Extraction strategy		Tasks		
			Patch	Region	Semantic(mIoU)↑	Depth(aErr) ↓	Normals(mErr)↓
full	TaskExpert* [44]	ViT-Large			57.58	0.3730	5.90
onelabel	Li *et al.* [25]	SegNet			30.36	0.6088	32.08
onelabel	DejaVu* [1]	SegNet			31.02	0.5959	32.15
onelabel	Gaussian Contrast(Ours)	SegNet		✓	31.79 (+2.48)	0.5835 (+2.25)	30.38 (+5.51)
onelabel	Li *et al.* * [25]	HRNet18			39.42	0.5071	14.79
onelabel	DejaVu [1]	HRNet18			—	—	—
onelabel	Gaussian Contrast	HRNet18	✓		41.56 (+5.43)	0.4884 (+3.69)	7.71 (+47.87)
onelabel	Gaussian Contrast(Ours)	HRNet18		✓	**42.28 (+7.26)**	**0.4641 (+8.48)**	**4.86 (+67.14)**
onelabel	TaskExpert* [44]	ViT-Large			49.53	0.4305	11.12
onelabel	TaskExpert [44]+Ours	ViT-Large		✓	55.81 (+12.68)	0.4092 (+4.95)	8.46 (+23.92)
random label	Li *et al.* [25]	SegNet			34.26	0.5787	31.06
random label	DejaVu* [1]	SegNet			35.72	0.5665	29.82
random label	Gaussian Contrast(Ours)	SegNet		✓	37.40 (+4.70)	0.5428 (+4.18)	28.97 (+2.85)
random label	Li *et al.* * [25]	HRNet18			41.35	0.4845	14.34
random label	DejaVu [1]	HRNet18			—	—	—
random label	Gaussian Contrast	HRNet18	✓		45.79 (+10.74)	0.4619 (+4.66)	7.37 (+48.61)
random label	Gaussian Contrast(Ours)	HRNet18		✓	**46.21 (+11.75)**	**0.4482 (+7.49)**	**4.49 (+68.69)**
random label	TaskExpert* [44]	ViT-Large			56.32	0.4282	6.47
random label	TaskExpert [44]+Ours	ViT-Large		✓	57.14 (+1.46)	0.3977 (+7.12)	4.78 (+26.12)

multiple tasks, with at least one task having a label and at most $K-1$ tasks having labels, where K represents the total number of tasks in the multi-task setup.

Implementation Details and Evaluation Metrics. All our experiments were conducted on the NVIDIA A100 GPU. For NYU-V2, we trained the models for 20 epochs, while for Cityscapes, we trained them for 100 epochs. Following [25, 29], we employed cross-entropy loss for semantic segmentation, L1 loss for depth estimation, and cosine similarity loss for surface normal estimation. We utilized the exact same evaluation metrics as mentioned in [25, 29]. For the semantic segmentation task, we employed the mean intersection over union (mIoU) metric. The absolute error (aErr) was used to evaluate the depth estimation task, and the mean error (mErr) was employed for surface normal estimation.

4.1 Quantitative Evaluation

Quantitative Results on NYU-V2 Dataset. The evaluation results are shown in Tab. 1. Due to the absence of pre-trained parameters in Li *et al.* [25], the training process exhibits slower convergence. To expedite the convergence speed, we replaced the backbone of Li *et al.* [25] with HRNet18, which possesses a smaller training parameter size but utilizes pre-trained weights. The comparisons between Li *et al.* [25] with SegNet as the backbone and Li *et al.* [25] with HRNet18 as the backbone are presented in the first row and the fourth row, respectively, for each label setting. Since the source code of DejaVu [1] is not publicly available, we are unable to replace its backbone with HRNet18 for a fair comparison. To resolve this issue, we provide the results of our method based

Table 2: Quantitative comparison on the Cityscapes dataset. ↑ (↓) denotes that, larger (smaller) values lead to better quality. * represents the available SOTA. + denotes the percentage improvement over the SOTA performance, corresponding to ↑ (↓). The bold denotes the best.

Label setting	Method	Backbone	Contrast strategy			Tasks	
			Vector	Pixel	Gaussian	Semantic(mIoU)↑	Depth(aErr) ↓
onelabel	Li *et al.* [25]	SegNet				74.90	0.0161
onelabel	Li *et al.* * [25]	HRNet18				81.73	0.0157
onelabel	Region-aware Contrast	HRNet18	✓			82.76 (+1.26)	0.0145 (+7.64)
onelabel	Region-aware Contrast	HRNet18		✓		83.20 (+1.80)	0.0141 (+10.19)
onelabel	Region-aware Contrast (Ours)	HRNet18			✓	**83.92 (+2.68)**	**0.0121 (+22.73)**

Table 3: Ablation study on the NYU-V2 dataset to explore the performance of different local extraction strategies and different local contrast strategies.

Label setting	Extraction strategy		Contrast strategy			Tasks		
	Patch	Region	Vector	Pixel	Gaussian	Semantic(mIoU)↑	Depth(aErr)↓	Normals(mErr)↓
onelabel	✓		✓			40.24	0.4998	7.9518
onelabel	✓			✓		41.50	0.4934	8.3648
onelabel	✓				✓	41.56	0.4884	7.7141
onelabel		✓	✓			41.84	0.4727	5.8098
onelabel		✓		✓		42.05	0.4663	4.9150
onelabel		✓			✓	**42.28**	**0.4641**	**4.8613**
random label	✓		✓			43.98	0.4795	7.4464
random label	✓			✓		44.62	0.4777	7.8873
random label	✓				✓	45.79	0.4619	7.3747
random label		✓	✓			45.83	0.4630	4.8393
random label		✓		✓		**46.41**	0.4518	4.8442
random label		✓			✓	46.21	**0.4482**	**4.4907**

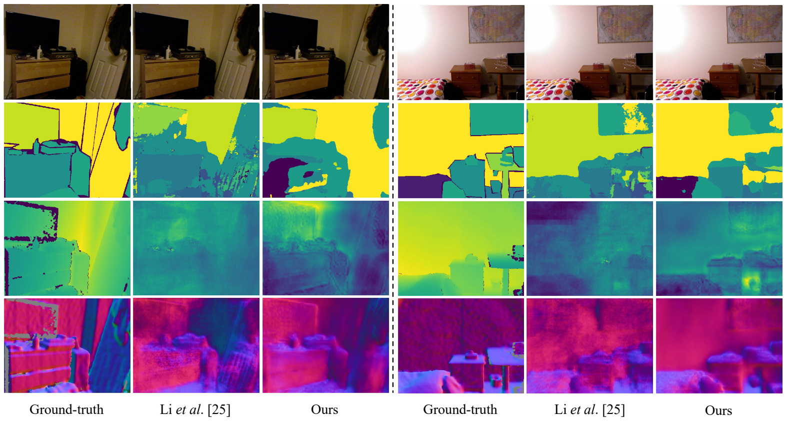


Fig. 4: Qualitative results of onelabel setting on NYU-V2. The first row shows the input image, the second row represents the ground-truth or predictions of semantic segmentation, the third row plots the ground-truth or predictions of depth estimation, and the final row presents the ground-truth or predictions of surface normal estimation.

on SegNet, which utilizes the same backbone. In all the methods compared using the same backbone, our approach demonstrates significant advantages. In the one label setting, the patch-aware Gaussian contrast method demonstrates

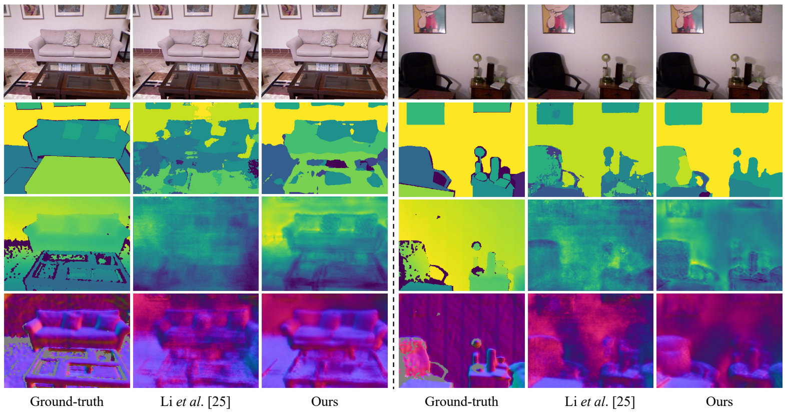


Fig. 5: Qualitative results of random label setting on NYU-V2. The first row shows the input image, the second row represents the ground-truth or predictions of semantic segmentation, the third row plots the ground-truth or predictions of depth estimation, and the final row presents the ground-truth or predictions of surface normal estimation.

significant improvements across all three tasks compared to Li *et al.* [25] across all three tasks, as it achieves a certain degree of local-level cross-task alignment. Furthermore, the region-aware Gaussian contrast method further enhances performance, achieving improvements of 2.48%, 2.25% and 5.51% over the SOTA [1] across the three tasks, owing to the achievement of cross-task consistency at the object level. In the random label setting, the superiority of the region-aware Gaussian contrast method can be observed, as it outperforms other methods significantly across all three tasks, yielding improvements of 4.70%, 4.18% and 2.85%, respectively, over the SOTA [1].

Furthermore, we provide a comparison with the state-of-the-art fully supervised method, TaskExpert [44]. Tab. 1 clearly shows that our method significantly improves the performance of TaskExpert in partially supervised scenarios. Particularly in the random setting, when combined with our method, TaskExpert achieves even better performance than the fully supervised TaskExpert in the surface normal task. This once again demonstrates the effectiveness and wide applicability of our approach.

Quantitative Results on Cityscapes Dataset. The evaluation results are describe in Tab. 2. Similar to NYU-V2, we first compared Li *et al.* [25] with HRNet18 as the backbone. Subsequently, we discussed the results of the region-aware contrast method under different contrast strategies. It is observed that Gaussian contrast demonstrates improvements in semantic segmentation and depth estimation compared to vector-based methods, indicating that representing regions using Gaussian distributions provides a more comprehensive and accurate representation than using vectors. In terms of computational speed, the

Table 4: Ablation study on the NYU-V2 dataset to investigate the contribution of SAM in the **fully supervised setting**.

Method	Vector	Gaussian	Semantic(mIoU)↑	Depth(aErr)↓	Normals(mErr)↓
Region-aware Contrast	✓		49.00	0.4206	8.1878
Region-aware Contrast (Ours)		✓	**51.14**	**0.3982**	**4.8108**

Gaussian contrast method is faster and more efficient, while still performing on par with pixel-based contrast methods.

4.2 Qualitative Evaluation

The qualitative evaluation of onelabel setting, random label setting on the NYU-V2 dataset, and onelabel setting on the Cityscapes dataset are shown in Fig. 4, Fig. 5 and Fig. 6, respectively. We present a comparative analysis of the results from two sets of images for each setting of each dataset. As indicated by Fig. 4 and Fig. 5, it can be observed that our method exhibits superior accuracy and smoothness in semantic segmentation, which is particularly evident in Fig. 5. Regarding depth estimation, our method, which is based on region-aware Gaussian distribution contrast, achieves more accurate predictions when objects undergo changes compared to Li *et al*. [25]. The latter fails to capture such variations effectively. As for surface normal estimation, our method outperforms Li *et al*. [25] overall and demonstrates finer contour handling. These observations highlight the advantages of our approach over Li *et al*. [25] in terms of classification accuracy, smoothness in semantic segmentation, accurate depth estimation in the presence of object changes, and finer handling of surface normals.

We further exhibit the multi-task prediction results on the Cityscapes dataset in Fig. 6. Benefited from our proposed region-aware distribution contrast learning method, both the semantic segmentation and depth estimation tasks exhibit improved capability in capturing object edges and fine-grained details compared to Li *et al*. [25]. Additionally, there are significant overall performance improvement at a global level.

4.3 Ablation Study

We conduct ablation study on NYU-V2 dataset to prove the effectiveness of our method from various aspects.

Local Extraction Strategy. Tab. 3 provides a comparison between different local extraction strategies. All methods are compared based on HRNet18 as the backbone. It can be observed that in both the onelabel setting and random label setting, regardless of the contrast strategy employed, the region-aware contrast method significantly outperforms the patch-aware contrast method. This is because the region-aware method takes into account the objects present in the image and performs contrast based on objects, rather than simply comparing patches that may contain multiple objects.

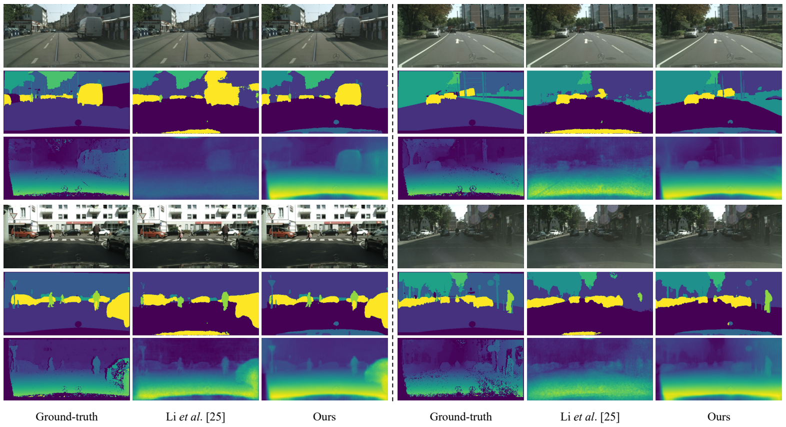


Fig. 6: Qualitative results of onelabel setting on Cityscapes. The first row shows the input image, the second row represents the ground-truth or predictions of semantic segmentation, and the final row plots the ground-truth or predictions of depth estimation.

Table 5: Ablation study on the NYU-V2 dataset to investigate the performance of various distance measurements.

Label setting	Method	Tasks		
		Semantic(mIoU)↑	Depth(aErr)↓	Normals(mErr)↓
onelabel	KL divergence	41.15	0.4739	4.94
onelabel	Wasserstein distance (Ours)	**42.28**	**0.4641**	**4.86**
random label	KL divergence	45.53	0.4539	5.04
random label	Wasserstein distance (Ours)	**46.21**	**0.4482**	**4.49**

Local Contrast Strategy. Tab. 3 also provides a comparison between different local contrast strategies. All methods are similarly compared based on HRNet18. It can be observed that regardless of the label setting and the local extraction strategy employed, the Gaussian distribution contrast consistently outperforms the vector contrast and approaches or even surpasses the pixel contrast. This is because multi-task learning scenarios involve discrete and continuous dense predictions, utilizing Gaussian distributions to represent regions for contrast allows for better cross-task alignment in the overall distribution. In contrast, the hard alignment strategy of pixel contrast may lead to inaccuracies.

Contribution of SAM. As shown in Tab. 2, Tab. 3, and Tab. 6 in our article, although the vector contrast approach also incorporates SAM, it fails to yield satisfactory results. In contrast, the adoption of distribution contrast exhibits a significant improvement over the vector contrast approach. To further demonstrate this point, we provide a comparison of results using semantic segmentation labels instead of SAM in fully supervised setting, as shown in Tab. 4. This further emphasizes that solely introducing fine-grained segmentation results is insufficient, highlighting the importance of the distribution contrast method.

Table 6: Ablation study on the NYU-V2 dataset to investigate the performance of various methods in the **fully supervised setting**.

Method	Contrast strategy			Tasks		
	Vector	Pixel	Gaussian	Semantic(mIoU)↑	Depth(aErr)↓	Normals(mErr)↓
MTL				46.38	0.4660	6.7298
Li *et al.* [25]				47.20	0.4477	5.1192
Region-aware Contrast	✓			47.64	0.4231	5.1072
Region-aware Contrast		✓		50.06	0.4168	4.5318
Region-aware Contrast (Ours)			✓	**50.83**	**0.4051**	**3.2878**

Distance Measurement for Distributions. Tab. 5 presents a comparison of different distance measurements. While KL divergence is commonly used to calculate the distance between Gaussian distributions, it fails to accurately reflect the distance between discrete and continuous distributions, which is crucial in the context of multi-task alignment. On the other hand, Wasserstein distance has the advantage of measuring the distance between two non-overlapping distributions. The experimental results in Tab. 5 demonstrate that Wasserstein distance is better suited for evaluating the distance between different task distributions. More detailed information can be found in the supplementary materials.

Fully Supervised Setting. Tab. 6 presents the comparison of our method in the fully supervised setting. All methods are based on HRNet18. Compared to other methods, the region-aware distribution contrast method once again demonstrates superiority, outperforming significantly in all three tasks. This is because Gaussian distributions can characterize the distribution characteristics of different tasks and facilitate accurate and efficient alignment with both vector and pixel-based approaches.

5 Conclusion

In this paper, we propose a novel region-aware distribution contrast learning method for MTPSL. To facilitate cross-task alignment at the region level, predictions from different tasks are initially mapped into a joint feature map space. Simultaneously, SAM is employed to generate regions for each training sample, leveraging its ability to provide valuable prior knowledge. Subsequently, region-wise features are modeled by fitting a Gaussian distribution. The alignment among regions is achieved by minimizing the disparity among distributions from different tasks within the same region and maximizing the distance between distributions from different regions. The effectiveness of our method is demonstrated on two standard multi-task datasets through extensive experiments, outperforming the state-of-the-art by a significant margin. Although our method is designed for MTPSL, it can be applied to general multi-task learning problems to leverage the intrinsic relationships across different tasks, as demonstrated in the fully supervised experiments in Sec. 4.3. Moreover, we believe that the proposed region-aware distribution contrast learning method provides a promising way for solving more general multi-modal problems, and we will investigate this in our future work.

Acknowledgements: This work was supported in part by the National Natural Science Foundation of China under grant U23B2011, 62102069, U20B2063 and 62220106008, the Key R&D Program of Zhejiang under grant 2024SSYS0091, the Sichuan Science and Technology Program under Grant 2024NSFTD0034.

References

1. Borse, S., Das, D., Park, H., Cai, H., Garrepalli, R., Porikli, F.: Dejavu: Conditional regenerative learning to enhance dense prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19466–19477 (2023)
2. Brüggemann, D., Kanakis, M., Obukhov, A., Georgoulis, S., Van Gool, L.: Exploring relational context for multi-task dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15869–15878 (2021)
3. Chang, J., Lan, Z., Cheng, C., Wei, Y.: Data uncertainty learning in face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5710–5719 (2020)
4. Chen, S., Xie, G., Liu, Y., Peng, Q., Sun, B., Li, H., You, X., Shao, L.: Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. Advances in Neural Information Processing Systems **34**, 16622–16634 (2021)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. pp. 1597–1607. PMLR (2020)
6. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
7. Chen, Z., Badrinarayanan, V., Lee, C.Y., Rabinovich, A.: GradNorm: Gradient normalization for adaptive loss balancing in deep multi-task networks. In: International Conference on Machine Learning. pp. 794–803. PMLR (2018)
8. Chen, Z., Zhu, L., Wan, L., Wang, S., Feng, W., Heng, P.A.: A multi-task mean teacher for semi-supervised shadow detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5611–5620 (2020)
9. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3213–3223 (2016)
10. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2650–2658 (2015)
11. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in Neural Information Processing Systems **27** (2014)
12. Fan, R., Poggi, M., Mattoccia, S.: Contrastive learning for depth prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3225–3236 (2023)
13. Gao, Y., Ma, J., Zhao, M., Liu, W., Yuille, A.L.: 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. pp. 3205–3214 (2019)

14. Guo, M., Haque, A., Huang, D.A., Yeung, S., Fei-Fei, L.: Dynamic task prioritization for multitask learning. In: Proceedings of the European Conference on Computer Vision. pp. 270–287 (2018)
15. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969 (2017)
17. Hu, H., Cui, J., Wang, L.: Region-aware contrastive learning for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16291–16301 (2021)
18. Huang, C., Tang, H., Fan, W., Xiao, Y., Hao, D., Qian, Z., Terzopoulos, D., et al.: Partly supervised multi-task learning. In: 2020 19th IEEE International Conference on Machine Learning and Applications. pp. 769–774. IEEE (2020)
19. Imran, A.A.Z., Terzopoulos, D.: Semi-supervised multi-task learning with chest x-ray images. In: Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10. pp. 151–159. Springer (2019)
20. Jin, X., Lan, C., Zeng, W., Chen, Z.: Global distance-distributions separation for unsupervised person re-identification. In: Proceedings of the European Conference on Computer Vision. pp. 735–751. Springer (2020)
21. Jin, Z., Chen, S., Chen, Y., Xu, Z., Feng, H.: Let segment anything help image dehaze. arXiv preprint arXiv:2306.15870 (2023)
22. Kendall, A., Gal, Y., Cipolla, R.: 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. pp. 7482–7491 (2018)
23. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (October 2023)
24. Latif, S., Rana, R., Khalifa, S., Jurdak, R., Epps, J., Schuller, B.W.: Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. IEEE Transactions on Affective Computing **13**(2), 992–1004 (2020)
25. Li, W.H., Liu, X., Bilen, H.: Learning multiple dense prediction tasks from partially annotated data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18879–18889 (2022)
26. Liang, C., Wang, W., Miao, J., Yang, Y.: Gmmseg: Gaussian mixture based generative semantic segmentation models. Advances in Neural Information Processing Systems **35**, 31360–31375 (2022)
27. Liu, Q., Liao, X., Carin, L.: Semi-supervised multitask learning. Advances in Neural Information Processing Systems **20** (2007)
28. Liu, S., James, S., Davison, A., Johns, E.: Auto-lambda: Disentangling dynamic task relationships. Transactions on Machine Learning Research (2022)
29. Liu, S., Johns, E., Davison, A.J.: End-to-end multi-task learning with attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1871–1880 (2019)
30. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015)

31. Lu, Y., Kumar, A., Zhai, S., Cheng, Y., Javidi, T., Feris, R.: Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5334–5343 (2017)
32. Misra, I., Shrivastava, A., Gupta, A., Hebert, M.: Cross-Stitch networks for multi-task learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3994–4003 (2016)
33. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
34. Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S.: On the uncertainty of self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3227–3237 (2020)
35. Rüschendorf, L.: The wasserstein distance and approximation theorems. Probability Theory and Related Fields **70**(1), 117–129 (1985)
36. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: Proceedings of the European Conference on Computer Vision. pp. 746–760. Springer (2012)
37. Standley, T., Zamir, A., Chen, D., Guibas, L., Malik, J., Savarese, S.: Which tasks should be learned together in multi-task learning? In: International Conference on Machine Learning. pp. 9120–9132. PMLR (2020)
38. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: Proceedings of the European Conference on Computer Vision. pp. 776–794. Springer (2020)
39. Vandenhende, S., Georgoulis, S., Van Gool, L.: MTI-Net: Multi-scale task interaction networks for multi-task learning. In: Proceedings of the European Conference on Computer Vision. pp. 527–543. Springer (2020)
40. Wang, X., Zhang, R., Shen, C., Kong, T., Li, L.: Dense contrastive learning for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3024–3033 (2021)
41. Wu, L., Zhong, Z., Fang, L., He, X., Liu, Q., Ma, J., Chen, H.: Sparsely annotated semantic segmentation with adaptive gaussian mixtures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15454–15464 (2023)
42. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3733–3742 (2018)
43. Xu, D., Ouyang, W., Wang, X., Sebe, N.: 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. pp. 675–684 (2018)
44. Ye, H., Xu, D.: Taskexpert: Dynamically assembling multi-task representations with memorial mixture-of-experts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21828–21837 (2023)
45. Zamir, A.R., Sax, A., Cheerla, N., Suri, R., Cao, Z., Malik, J., Guibas, L.J.: Robust learning through cross-task consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11197–11206 (2020)
46. Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3712–3722 (2018)
47. Zhang, J., Fan, D.P., Dai, Y., Anwar, S., Saleh, F.S., Zhang, T., Barnes, N.: Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoen-

coders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8582–8591 (2020)

48. Zhang, L., Liu, X., Guan, H.: Automtl: A programming framework for automating efficient multi-task learning. Advances in Neural Information Processing Systems **35**, 34216–34228 (2022)
49. Zhang, M., Zhao, X., Yao, J., Yuan, C., Huang, W.: When noisy labels meet long tail dilemmas: A representation calibration method. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15890–15900 (2023)
50. Zhang, Z., Cui, Z., Xu, C., Yan, Y., Sebe, N., Yang, J.: Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4106–4115 (2019)