Contrastive Multi-Task Dense Prediction

Siwei Yang[1,2], **Hanrong Ye**[2], **Dan Xu**[2]

[1] Key Laboratory of Embedded System and Service Computing, Tongji University
[2] Hong Kong University of Science and Technology
swyang.ac@gmail.com, hyeae@cse.ust.hk, danxu@cse.ust.hk

Abstract

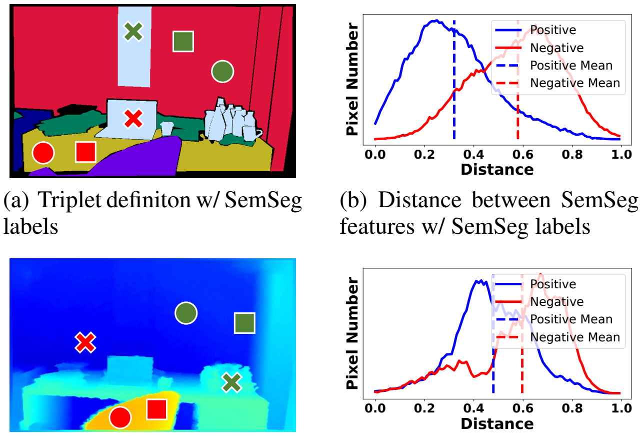
This paper targets the problem of multi-task dense prediction which aims to achieve simultaneous learning and inference on a bunch of multiple dense prediction tasks in a single framework. A core objective in design is how to effectively model cross-task interactions to achieve a comprehensive improvement on different tasks based on their inherent complementarity and consistency. Existing works typically design extra expensive distillation modules to perform explicit interaction computations among different task-specific features in both training and inference, bringing difficulty in adaptation for different task sets, and reducing efficiency due to clearly increased size of multi-task models. In contrast, we introduce feature-wise contrastive consistency into modeling the cross-task interactions for multi-task dense prediction. We propose a novel multi-task contrastive regularization method based on the consistency to effectively boost the representation learning of the different sub-tasks, which can also be easily generalized to different multi-task dense prediction frameworks, and costs no additional computation in the inference. Extensive experiments on two challenging datasets (*i.e.* NYUD-v2 and Pascal-Context) clearly demonstrate the superiority of the proposed multi-task contrastive learning approach for dense predictions, establishing new state-of-the-art performances.

Introduction

Dense prediction tasks such as semantic segmentation, depth estimation, and saliency detection empowered with deep learning techniques are moving very rapidly in the recent years, and methods with convolutional neural networks (CNNs) have demonstrated great improvement on these different tasks. However, most of the methods usually address these dense prediction tasks separately which results in low efficiency in real-life applications utilizing large-capacity CNNs. More importantly, simultaneous modeling of multiple different tasks allows us to capture the interior relationships and interactions among the different tasks, being able to realize a more powerful and higher-level perception system.

Multi-task dense prediction (Kanakis et al. 2020; Vandenhende et al. 2020; Brüggemann et al. 2021) offers an effective solution to this problem by jointly learning both task-sharing and task-specific representations. As the task-sharing representations can be obtained from a common backbone, it greatly facilitates the reduction of memory and computation overhead in both training and testing. However, simply using a shared backbone and several individual decoding heads for different tasks often causes a clear performance drop compared to a single-task dense prediction framework (Xu et al. 2018). Therefore, how to design a deep network structure that can effectively model cross-task interactions based on the complementarity of different tasks is critical for multi-task dense prediction.

To achieve the above-mentioned objective, recent works on multi-task dense prediction mainly employ two paradigms. One is to refine final task-specific features with a multi-modal distillation module via carefully designed attention mechanisms to improve the final prediction (Xu et al. 2018; Vandenhende et al. 2020; Brüggemann et al. 2021).



(a) Triplet definiton w/ SemSeg labels

(b) Distance between SemSeg features w/ SemSeg labels

(c) Triplet definiton w/ Depth labels

(d) Distance between SemSeg features w/ Depth labels

Figure 1: The feature distance distributions of positive/negative pairs show high cross-task consistency, which can be utilized to boost multi-task pixel-wise learning and predictions. In (a) and (c), anchor pixels are denoted with □, positive pixels with ∘, and negative pixels with ×. (b) and (d) illustrates the feature distributions of the positive/negative pixel pairs in the projection feature space.

The other is to learn and combine intra- or inter-task affinity maps to refine the task-specific features (Zhang et al. 2019; Zhou et al. 2020). A common point for these two categories of methods is that they need to have extra expensive network computation modules for cross-modal or cross-task interactions, especially when the computation is performed in a multi-scale setting, which will significantly increase the training and inference cost.

In this work, in contrast to existing methods focusing on network structure improvement, we present a novel approach for the targeted problem based on learning cross-task contrastive consistency regularization upon dense pixel-wise features. The intuition of of utilizing contrastive consistency for modeling cross-task interactions is mainly threefold. First, as we learn the different tasks from the same input image data, the consistency inherently exists among the different tasks (see Fig. 1). Second, the features from different task decoders corresponding to the same semantic object categories should be more similar and consistent in the feature space than those with different object categories. Third, the extra network modules for explicit cross-task distillation in previous works inevitably increase the complexity of the multi-task model, bringing larger computation overhead in both the training and testing phases. However, a contrastive consistency optimization objective on multi-task features would only bring computation in the training stage, while the model size and the testing efficiency can be effectively improved.

Based on these motivations, we develop a contrastive learning approach for multi-task dense prediction, which leverages cross-task consistency and applies contrastive regularization onto the features of different tasks. The feature-level multi-task contrastive optimization objective can guide the model to learn more effective task-specific features via absorbing complementary information from other tasks, without an explicit utilization of extra cross-task interaction network structures. The proposed approach can also be flexibly generalized, in a plug and play fashion, to other multi-task dense prediction frameworks that may have different network architectures and objective functions for discrete or continuous dense prediction tasks.

The main contribution of this paper is threefold:

- We demonstrate pixel-level feature consistency exists among multiple distinct dense prediction tasks and introduce feature-wise contrastive consistency to guide the learning of discriminative multi-task dense features, and propose a novel cross-task contrastive learning strategy for the problem of joint multi-task dense predictions.
- We implement an end-to-end multi-task contrastive regularization framework based on the feature consistency, and further design effective schemes including a generic criterion for positive/negative definition for both continuous and discrete tasks, shared feature projection, semi-hard pixel sampling, and contrastive task-pair selection to advance the framework.
- Extensive experiments on two challenging datasets (*i.e.* NYUD-v2 and Pascal-Context) clearly demonstrate the effectiveness of the proposed cross-task contrastive regularization model for multi-task dense prediction, establishing new state-of-the-art performance. The results also verify that the proposed model can be generalized to different existing multi-task dense prediction frameworks to boost the performance and costs no additional computation in inference.

Related Work

We review closely related works on multi-task dense prediction and contrastive learning for vision problems.

Multi-task Dense Prediction

Recent methods on multi-task dense prediction can be roughly divided into two groups: the first group of methods (Kendall, Gal, and Cipolla 2018; Chen et al. 2018; Neven et al. 2017; Sener and Koltun 2018; Teichmann et al. 2018; Gao et al. 2019; Liu, Johns, and Davison 2019) shares task-specific information in the encoding stage while the other group of methods shares the information in the decoding stage (Xu et al. 2017; Brüggemann et al. 2021; Zhang et al. 2019; Zhou et al. 2020). We focus on the review of the second group of methods since their setting is closer to ours.

One common practice among decoder-focused methods is to use features from other tasks to refine the final features or predictions via cross-task or cross-modal distillation (Xu et al. 2018; Vandenhende et al. 2020; Brüggemann et al. 2021). More specifically, PAD-Net (Xu et al. 2018) applies a multi-task distillation with a spatial attention mechanism to enhance task-specific predictions while at a fixed scale, leading to sub-optimal performance. To alleviate this issue, MTI-Net (Vandenhende et al. 2020) proposes to apply multi-modal distillation at different scales in a parallel way. Despite distillation at different scales brings diverse receptive fields, the information interaction crossing tasks is still restricted in a local context. Thus, Bruggemann *et al.* (Brüggemann et al. 2021) design an ATRC module to leverage both local and global contexts, based on not only task-specific features as in PAD-Net and MTI-Net but also task-specific predictions in a dynamic and adaptive manner. Another research line of decoder-focused models, for instance, PAP (Zhang et al. 2019) and PSD (Zhou et al. 2020), aims to refine task-specific feature maps by mining local or global affinities on task-specific features.

In contrast to these two groups of methods which design extra network modules to explicitly perform cross-task interaction and refine features during inference, our method directly regularizes the model learning to produce more effective multi-task features during training, using a novel multi-task contrastive learning approach. Thus, our model can simplify the model complexity, and can also be more flexibly applied in a plug and play way into other multi-task dense prediction frameworks.

Contrastive Learning for Vision

Contrastive learning has been widely used to deal with various fundamental representation learning and its application to downstream single tasks problems (Van den Oord, Li, and Vinyals 2018; He et al. 2020; Chen et al. 2020c,a,b).

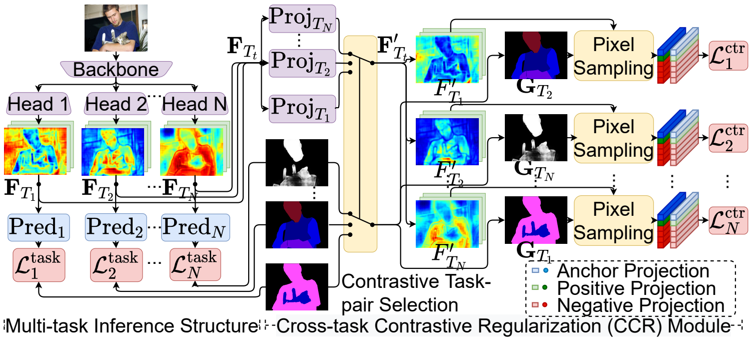


Figure 2: At each training iteration, a random source task is selected for every target task independently, whose ground-truth label $\mathbf{G}_{T_s}$ is used to guide the regularization applied on target-task feature map $\mathbf{F}_{T_t}$. A projector Proj$_{T_t}$ for every pair of target and source tasks produces a feature projection map $\mathbf{F}'_{T_t}$. A sampling strategy guided with positive/negative definition based on $\mathbf{G}_{T_s}$ is applied on projection maps, and the selected pixels are then used for the triplet loss. Best viewed in color.

One challenging *single* dense prediction task, *i.e.* semantic segmentation, has greatly benefited from the contrastive learning philosophy in its semantic representation learning (Hu, Cui, and Wang 2021; Alonso et al. 2021; Zhao et al. 2021; Zhong et al. 2021; Chaitanya et al. 2020; Van Gansbeke et al. 2021). For instance, RegionContrast (Hu, Cui, and Wang 2021) applies contrastive regularization on features to enhance the similarity between pixels corresponding to the same category, while Zhao *et al.* (Zhao et al. 2021) adopt contrastive learning as a pretraining strategy when training data is insufficient. Some other works leverage contrastive learning for semantic segmentation in a self-supervised (Zhang et al. 2021; Van Gansbeke et al. 2021) or semi-supervised (Chaitanya et al. 2020; Zhong et al. 2021; Alonso et al. 2021) manner.

However, in contrast to these works considering learning only with a *single* dense prediction task, our work aims at improving the comprehensive performance of multi-task dense predictions involving multiple distinct tasks via cross-task contrastive consistency.

Contrastive Multi-Task Dense Prediction

A framework overview of the proposed approach is shown in Fig. 2. The proposed contrastive multi-task regularization model is constructed based on dense feature consistency, and is applied upon different task-specific feature maps from the multiple decoding heads, to boost their representation learning. It consists of several designed important components, including shared feature projectors, sampling strategies of spatial feature points of different tasks, the definition criterion of positive and negative samples, and the multi-task contrastive learning objective. In the following, we first discuss the cross-task feature consistency, and then elaborate the details of the proposed model.

Cross-Task Dense Feature Consistency

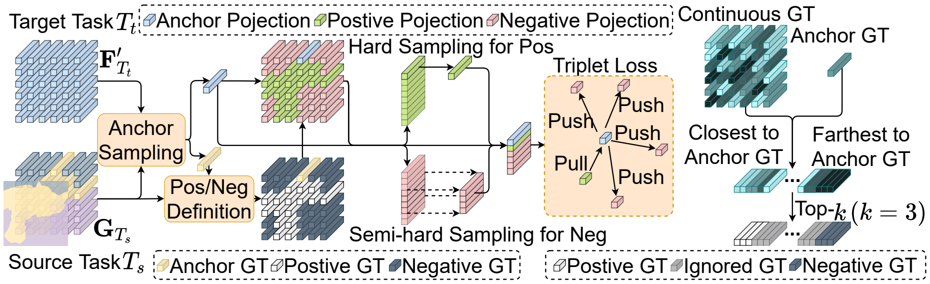
From a basic principle of image-based representation learning, the pixels with the same ground-truth labels distribute more closely (*i.e.* with smaller distance) in the feature space compared to those with different ground-truth labels. In the context of multi-task dense prediction, as the different tasks are learned from the same input image data, consistency inherently exists crossing the different tasks. The model performance of one task can thus benefit from utilizing the consistent information from other tasks. To better illustrate such consistency, we also train the multi-task baseline model on NYUD-v2 (Silberman et al. 2012), and sample 10^5 pixel triplets, each with one anchor pixel, one positive pixel and one negative pixel, from the feature maps that are used for final predictions. Detailed criterion of positive or negative sample definition and the training setting are elaborated in later sections. The distance is calculated with a squared L2 between two normalized pixel features. As shown in Fig. 1, the feature distance distribution shows high consistency on different target tasks, *e.g.*, distances of positive pairs are consistently smaller than those of negative ones on different target tasks, and the distance distributions are also similar. Based on such observations, we design a multi-task contrastive regularization model to utilize the cross-task consistency to enhance the multi-task representation learning.

Multi-Task Contrastive Regularization

We first introduce the definition criterion of pixel-wise positive and negative samples, and then present the details of the proposed pixel-wise feature contrastive regularization based on cross-task consistency for task-specific feature learning, and important schemes to boost the learning performance, including the shared feature projector, the semi-hard pixel sampling strategy, and the contrastive task-pair selection.

Positive/Negative Sample Definition Dense prediction tasks can be divided into two types according to the continuity of the ground-truth label space, *i.e.* pixel-wise classification tasks with *discrete* ground-truths, *e.g.*, semantic segmentation and human-parts parsing, and pixel-wise regression tasks with *continuous* ground-truths, *e.g.*, depth and surface normal estimation. For the proposed multi-task dense contrastive learning objective, we introduce two general definition criteria for positive and negative samples for these two different types of tasks:

(i) Tasks with discrete ground-truth labels. It is relatively



(a) Pixel Sampling in Cross-task Contrastive Regularization (b) Pos/Neg Definition on Continous GT

Figure 3: (a) A detailed illustration of the designed semi-hard pixel sampling process for the proposed Cross-Task Contrastive Regularization (CCR) method. Anchor pixels are first sampled, whose corresponding labels are then compared with the labels of other pixels to determine whether the projection of a certain pixel should be regarded as a positive or a negative sample. Hard sampling is applied to positive ones, while semi-hard sampling is applied to negative ones. (b) A generic positive/negative definition of continuous tasks.

straightforward to define positive and negative pixel samples as the pixels with the same ground-truth labels can be considered as positives while those with different labels can be considered as negatives. We simply apply this rule to determine positive and negative pixel samples for this group of dense prediction tasks.

(ii) Tasks with continuous ground-truth labels. Different from those tasks with discrete ground-truth labels, there is no clear boundary to distinguish positive and negative samples for tasks with continuous ground-truth labels. Thus, for a certain anchor pixel that is randomly selected, we propose to utilize top-k pixels with closer distance to the anchor pixel as positive samples, and top-k pixels with farther distance to the anchor pixel as negative ones. Since different evaluation metrics for dense prediction tasks are usually based on pixel-wise comparison, the measurement of pixel-wise distances in the label space can be based on the evaluation metric for the corresponding task, *e.g.*, the L1 distance for depth estimation, and the angular error for surface normal estimation.

With the proposed definition criteria, we can ensure that our model can be flexibly generalized to any continuous or discrete dense prediction tasks.

Pixel-wise Feature Contrastive Regularization We now present the proposed cross-task contrastive regularization (CCR) for discriminative multi-task feature learning. Given N different dense predictions tasks, any task can be a source task T_s ($s \in \{1, \dots, N\}$ or a target task T_t ($t \in \{1, \dots, N\}$). To enforce the cross-task contrastive learning, C_A anchor positions are first sampled from a target-task-specific feature map $\mathbf{F}_{T_t}$. The sampling in the basic CCR module considers a uniform sampling strategy. The ground-truth labels $\mathbf{G}_{T_s}$ of a source task T_s are used to define the triplets based on the positive/negative sample definition presented in the previous section, and we generate a set of C_A triplets $\{a^{c-}_{T_s}, a^{c}_{T_s}, a^{c+}_{T_s}\}^{C_A}_{c=1}$, where $a^{c+}_{T_s}$ and $a^{c-}_{T_s}$ are positive and negative pixels of an anchor pixel $a^{c}_{T_s}$, respectively. Then, a cross-task contrastive regularization $\mathcal{R}^{\text{ctr}}_{ts}$ is defined as:

$$\mathcal{R}^{\text{ctr}}_{ts} = \frac{1}{C_A}\sum_{c=1}^{C_A}\Big[D\big(\mathbf{F}'_{T_t}(a^{c}_{T_s}), \mathbf{F}'_{T_t}(a^{c+}_{T_s})\big) - D\big(\mathbf{F}'_{T_t}(a^{c}_{T_s}), \mathbf{F}'_{T_t}(a^{c-}_{T_s})\big) + m\Big]_+, \tag{1}$$

where $D(\cdot)$ is a distance measurement function, and an L2 distance is utilized for simplicity. $\mathbf{F}'_{T_t}$ is a normalized feature map produced from a task-specific feature projector $\text{Proj}_{T_t}(\cdot)$ as $\mathbf{F}'_{T_t} = \text{Proj}_{T_t}(\mathbf{F}_{T_t})$. It can be represented as:

$$\text{Proj}_{T_t}(\cdot) = \text{BN} \circ \text{Conv}_{1\times1} \circ \text{ReLU} \circ \text{Conv}_{1\times1}(\cdot) \tag{2}$$

where BN is a batch normalization; $\text{Conv}_{1\times1}$ is a convolution with 1x1 kernel; ReLU is the ReLU activation function. $\mathbf{F}_{T_t}(\cdot)$ returns a feature vector given an input pixel location on the feature map. It should be noted that the ground-truth maps are downsampled to match the resolution of the projected feature map $\mathbf{F}_{T_t}$ for the contrastive loss computation. m is a margin variable that is enforced between positive and negative feature points. If we utilize the contrastive regularization on every pair of a source task T_s and a target task T_t, the overall multi-task contrastive regularization term writes:

$$\mathcal{L}^{\text{ctr}} = \sum_{t=1}^{N}\sum_{s=1, s\neq t}^{N}\mathcal{R}^{\text{ctr}}_{ts}\big(\mathbf{F}_{T_t}, \mathbf{G}_{T_s}, C_A, m\big). \tag{3}$$

The computation of the above-proposed multi-task contrastive learning objective can be implemented as an end-to-end learnable module, *i.e.* the proposed Cross-task Contrastive Regularization (CCR) module as illustrated in Fig. 2. A semi-hard pixel sampling scheme produces better sampled positive and negative feature points, and a contrastive task-pair selection determines which pair of tasks to be chosen for constructing an efficient regularization. We elaborate the details of these two components in the next parts.

Semi-hard Pixel Sampling Pixel features are considered as independent samples in our approach since the proposed multi-task contrastive regularization is applied in a pixel level. It is obvious that using all the pixels in every image requires very high computational overhead, and more importantly, not every pixel delivers critical information for the learning. Thus, we propose a pixel sampling strategy to address these issues. It follows several steps: **(i)** A certain amount of anchor pixels are uniformly sampled with a fixed ratio γ to the total number of pixels on an image $\mathbf{I} \in \mathbb{R}^{H \times W}$ downsampled to the resolution of projection map p_t, where H and W indicate the height and width of the image, respectively. This process produces $C_A = \gamma HW$ anchors, *i.e.* $a^c_{T_s} \sim U(\mathbf{I})$. **(ii)** With the ground truth $\mathbf{G}_{T_s}$ from a source task T_s, for each anchor pixel $a^c_{T_s}$, all the remaining pixels in $\mathbf{I}$ unless ignored are divided into a positive pixel set $\mathbf{I}^+_{T_s}$ or a negative pixel set $\mathbf{I}^-_{T_s}$. **(iii)** To sample positive and negative pixels, the L2 distances between anchors and all the pixels in the projected feature space are calculated, resulting in a distance matrix $M \in \mathbb{R}^{C_A \times HW}$. To achieve a faster convergence, hard pixels should have higher priority to be sampled. **(iv)** For every anchor pixel $a^c_{T_s}$, the most distant positive pixel is selected as its only positive sample:

$$a^{c+}_{T_s} = \underset{a \in I^+_{T_s}}{\arg\max}\, D\left(\mathbf{F}'_{T_t}(a^c_{T_s}), \mathbf{F}'_{T_t}(a)\right). \tag{4}$$

It is regarded as the hardest positive sample to be learned. **(v)** We also use semi-hard mining (Schroff, Kalenichenko, and Philbin 2015) to sample negative pixels to prevent the feature projector $\text{Proj}_{T_t}(\cdot)$ from collapsing. For every anchor pixel $a^c_{T_s}$, a set of C_{neg} negative pixels is selected, with each negative pixel $a^{c-}_{T_s}$ having the smallest distance with $a^c_{T_s}$ while still satisfying $\text{D}(\mathbf{F}'_{T_t}(a^c_{T_s}), \mathbf{F}'_{T_t}(a^{c-}_{T_s})) > \text{D}(\mathbf{F}'_{T_t}(a^c_{T_s}), \mathbf{F}'_{T_t}(a^{c+}_{T_s}))$. The proposed pixel sampling scheme is applied on each image independently to guarantee a large diversity of positive or negative pixel samples.

Shared Feature Projector Feature projection is critically important in the contrastive learning. Since the contrastive regularization in our model is applied on different combinations of target and source tasks, in which each task provides feature map and ground-truth labels for the cross-task regularization. In other words, any task can be a source task or a target task, and the source task provides labels to regularize the feature map of the target task through the positive/negative sample selection. Considering multiple tasks are simultaneously learned, if we have all projectors unshared, the projector capacity for all the tasks is very large, which may increase the difficulty of learning. Thus, we design different sharing strategies for learning the projectors for the different source and target tasks. If all the target tasks use the same feature projector, we notate this projector design as 'target-task shared (T_t-shared)'. If a target task uses the same projector for any paired source tasks, which provide distinct ground truth label for the cross-task regularization, we refer this projector as 'source-task shared (T_s-shared)'. The proposed sharing strategy can help not only achieve the learning efficiency but also improve the performance. Detailed results and analysis can be found in the experiments.

Contrastive Task-pair Selection Each cross-task constrastive regularization is constrained on a pair of tasks. When the number of tasks (*i.e.* N) is very large, we have $N \times (N - 1)$ possible combinations, which brings tremendously large-scale contrastive computation in the learning process. To simplify and make a reasonable optimization space, we propose a contrastive task-pair selection strategy to address this issue. Specifically, in each iteration of the optimization, for each target task T_t ($t \in \{1, ..., N\}$), we randomly select one source task from the rest $N - 1$ tasks. Since we perform random sampling, for each iteration, the same target task may pick up a different source task. After the whole training procedure, each target task can approximately obtain contrastive regularization from any source task in the whole task set.

Model Implementation Details

Overall Learning Objective The overall multi-task learning objective $\mathcal{L}_{overall}$ can be written as:

$$\mathcal{L}_{overall} = \sum_{i=1}^{N} \lambda^{\text{task}}_i \mathcal{L}^{\text{task}}_i + \lambda^{\text{ctr}} \mathcal{L}^{\text{ctr}}, \tag{5}$$

where $\mathcal{L}^{\text{task}}_i$ is a task-specific optimization loss for task T_i (*i.e.* a classification or a regression loss), and λ^{task}_i denotes a loss weight for each task. λ^{ctr} is the loss weight for all the cross-task contrastive losses. The λ^{ctr} is linearly ramped up in the first several epochs to stabilize the training.

Model Inference In the inference stage, the cross-task contrastive learning module can be removed from the whole multi-task framework for the inference of different tasks, which is also a significant advantage compared to previous multi-task dense prediction frameworks that design extra expensive distillation modules for the cross-task interaction, and also needs to involve the modules in the inference stage.

Experiments

Experimental Setup

Datasets The experiments are extensively conducted on two widely used multi-task dense prediction datasets. One is NYUD-v2 (Silberman et al. 2012) which contains 1,449 RGBD indoor scene images with annotations for tasks of semantic segmentation (SemSeg), depth estimation (Depth), and surface normal estimation (Normal), with 795 images for training and 654 images for testing. The other one is Pascal-Context (Everingham et al. 2010) which has 4,998 training and 5,105 testing images labeled for tasks of semantic segmentation (SemSeg), human-parts parsing (Parsing), saliency estimation (Saliency), surface normal estimation (Normal), and edge detection (Edge).

Evaluation Metrics Following previous works (Vandenhende et al. 2020), SemSeg and Parsing tasks are are evaluated with mean Intersection over Union (mIoU), Depth with root mean square error (rmse), Normal with mean angular error (mErr), Salience with maximum F_1 score (maxF1), and Edge with optimal-dataset-scale F-measure (odsF). The overall multi-task performance introduced in (Maninis, Radosavovic, and Kokkinos 2019) is measured with an average of per-task performance differences w.r.t. a corresponding single-task baseline trained separately.

Training Settings Similar to MTI-Net (Vandenhende et al. 2020), our ablation studies are extensively conducted on NYUD-v2 and the HRNet18 (Wang et al. 2020) is used as the model backbone which is pretrained on ImageNet (Deng

Model	SemSeg↑	Depth↓	Normal↓	$\Delta_m(\%)\uparrow$
ST Baseline	39.803	0.617	19.896	-
MT Baseline	38.901	0.615	20.712	-2.025
CCR-Basic	39.499	0.609	20.305	0.184
+ Proj	39.956	0.603	20.297	0.635
+ Proj + SS	40.779	0.595	20.240	1.203
+ Proj + SS + CTS	**41.275**	**0.592**	**20.178**	**2.101**

Table 1: Overall ablation study to show the effectiveness of the proposed cross-task contrastive regularization (CCR).

Model	SemSeg↑	Depth↓	Normal↓
PAD-NET	35.406	0.670	21.991
PAD-NET + CCR	**36.925**	**0.654**	**21.139**
MTI-NET	37.456	0.626	21.031
MTI-NET + CCR	**39.192**	**0.607**	**20.826**
InvPT	52.840	0.514	18.872
InvPT + CCR	**53.799**	**0.508**	**18.670**

Table 2: Performance improvements with the proposed CCR applied on different best-performing models on NYUD-v2.

Model	T_t	T_s	SemSeg↑	Depth↓	Normal↓
MT Baseline			38.901	0.615	20.712
CCR			40.414	0.598	20.219
CCR	✓		40.273	0.604	20.318
CCR	✓	✓	40.574	0.600	20.299
CCR		✓	**41.275**	**0.592**	**20.178**

Table 3: Ablation study on feature projector designs. T_t/T_s means projectors are shared among different target/source tasks. SS and CTS are included in all CCR models.

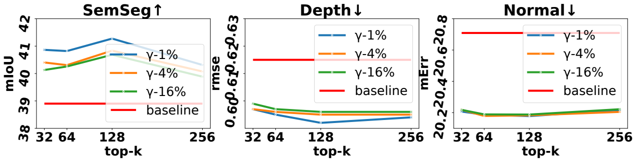


Figure 4: Investigation on top-k and sampling ratio γ. It can be observed from the performance variance that the model is not very sensitive to k and γ, while k has a greater influence on model performance than γ does. The best performance is achieved when k is set to 128.

et al. 2009). We train each model using Adam optimizer with a batch size of 4 on 2 GPUs (*i.e.* NVIDIA RTX 3090) for the NYUD-v2 dataset, and a batch size of 6 on 6 GPUs for the PASCAL-Context dataset. The base learning rate, momentum, and weight decay are set to 2e-4, 0.9, and 1e-4, respectively. The learning rate is linearly warmed up for 1 epoch. The margin m, sampling ratio γ, top-k factor k, and contrastive loss weight λ^{con} are by default set to 0.2, 0.01, 128, and 1.0 respectively in the evaluation experiments. The number of negative samples in one triplet is 16.

Model Analysis

Model Baselines and Variants Different baselines and model variants are considered for the evaluation: (i) **'ST Baseline'** is a very strong single-task baseline model using a shared FPN encoder (Lin et al. 2017) with the HRNet18 backbone and a SemanticFPN decoder (Kirillov et al. 2019), which is similar to the baseline widely used in existing state-of-the-art models (Vandenhende et al. 2020), while with better performance. (ii) **'MT Baseline'** indicates a multi-task baseline, which uses the same encoder architecture as the 'ST Baseline', while it has task-specific decoders for the different tasks which are jointly optimized with the encoder under the multi-task learning setting. (iii) **'CCR-Basic'** directly applies the proposed contrastive regularization on the final multi-task feature maps, based on the 'MT Baseline'. (iv) **'+ Proj'** indicates applying a shared feature projector on the multi-task features before we feed them into the CCR module. (v) **'+ SS'** means using the designed semi-hard pixel sampling. (vi) **'+ CTS'** denotes using the proposed contrastive task-pair selection strategy.

Overall Ablation Study To validate the effectiveness of our approach, an overall ablation study is shown in Table 1. Our basic CCR module (*i.e.* CCR-Basic) can bring notable performance gain over the baseline model (*i.e.* MT Baseline) on all the tasks, especially on SemSeg and Depth. The addition of the proposed shared projector strategy (*i.e.* Proj) further improves the performance on all the tasks. More experiments about the projector design are discussed later in this section. The semi-hard pixel sampling (*i.e.* SS) can also clearly boost the model's multi-task performance. Further more, it can be seen that the the contrastive task-pair selection strategy (*i.e.* CTS) which is originally designed to reduce the memory consumption, can also contribute to the model's performance. This may be due to that choosing one source task for each target task instead of pairing all the other tasks during each iteration can alleviate the gradient conflict issue on learning the shared source task projector.

Study of Projector Designs We study different projector designs and show the results in Table 3. It can be observed that how projectors are shared among different constrastive task pairs can greatly affect the performance. We find that sharing projectors among source tasks achieves the best performance, which indicates that the cross-task contrastive regularization should be applied in a joint feature projection space regardless of which source task is used for the definition of positive and negative pairs, while sharing projectors among different target tasks leads to less optimal performance. This is because that the feature distributions of different target tasks vary largely, and we need different projectors for the distinct target tasks to learn the joint feature projection space for the multi-task contrastive learning.

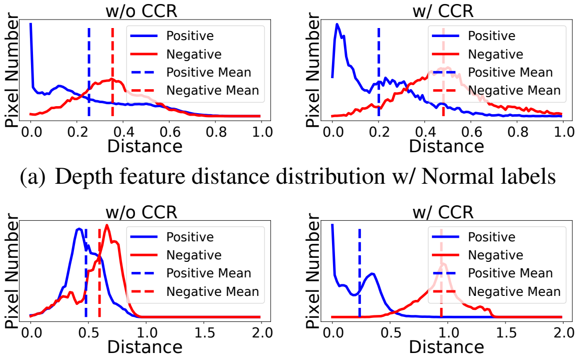
Study of Top-k Factor and Sampling Ratio γ Since there are no discriminative boundaries between positive and negative samples based on continuous ground-truth labels, we introduce top-k-based method to define positive/negative samples as illustrated in the method section. If k is too small, little contrastive constraint will be provided as sampled pos-

Model	SemSeg↑	Depth↓	Normal↓
PAP	36.72	0.618	20.82
PSD	36.69	0.625	20.87
ATRC	46.33	0.536	20.18
InvPT	52.84	0.514	18.87
CCR	**53.80**	**0.508**	**18.67**

Table 4: State-of-the-art comparison on NYUD-v2.

Model	SemSeg↑	Parsing↑	Saliency↑	Normal↓	Edge↑
PAD-NET	53.60	59.60	65.80	15.30	72.50
MTI-NET	61.70	60.18	84.78	14.23	70.80
PSD	61.70	60.18	84.78	14.23	70.80
ATRC	67.67	62.93	82.29	14.24	72.42
InvPT	79.03	67.61	**84.81**	14.15	73.00
CCR	**80.23**	**68.40**	84.36	**13.93**	**73.07**

Table 5: State-of-the-art comparison on Pascal-Context.



(a) Depth feature distance distribution w/ Normal labels

(b) SemSeg feature distance distribution w/ Depth labels

Figure 5: Feature distance distributions of feature maps of different tasks, *i.e.* (a) Depth and (b) SemSeg, based on sampled positive/negative pixel pairs defined using the different source task labels, *i.e.* (a) Normal and (b) Depth. Distance in the projection feature space is measured with normalized $l2$ distance. After using the proposed CCR, the average distance of the positive and negative features are significantly pushed apart, achieving more discriminative features.

itive and negative pixels can be easily discriminated with the triplet contrastive loss, while a very large k also brings ambiguity in the sample definition, and even causes intersection between positive and negative pixel samples, resulting in a learning corruption of the model. Figure 4 shows an evaluation of the influence of the factor k on the performance of different tasks. The best multi-task performance is achieved when k is set to 128, and the performance gets consistently improved as the sampling ratio γ increases, while the model performance is not very sensitive to the k and γ.

Study of Improvements on Different Baseline Models A simple yet effective regularization approach should be independent of the baseline models utilized. In Table 2, we demonstrate the performance of our approach when applied upon the most advanced multi-task dense predictions models, including PAD-Net (Xu et al. 2018), MTI-Net (Vandenhende et al. 2020), and InvPT (Ye and Xu 2022). It is clear that our approach can effectively boost the performances of all these baseline models, demonstrating the generalization ability of our model for multi-task dense prediction.

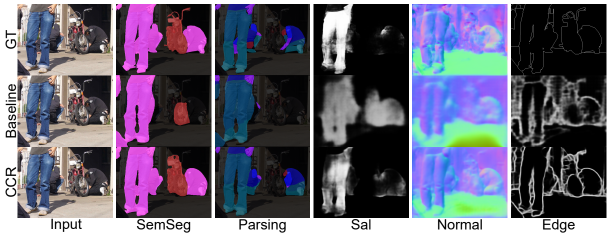


Figure 6: Qualitative comparison with the results of the baseline model and the ground-truths on Pascal-Context. CCR yields more accurate predictions on all the tasks.

Qualitative Study of the Effect on Feature Distributions To study the effect of the proposed CCR on the feature learning, we show the distance distributions of the pixel-wise feature triplets in the projection feature space, for both the baseline and our model in Fig 5. It can be observed that, with the proposed CCR, the pixel-wise features of different labels can be further pushed apart in the feature space, verifying our motivation of using the cross-task contrastive consistency for discriminative multi-task feature learning.

State-of-the-art Comparison

We compare the proposed contrastive multi-task learning approach with the single-task baseline, the multi-task baseline, and several best performing state-of-the-art multi-task dense prediction models in the literature, including PAD-Net (Xu et al. 2018), PAP (Vandenhende et al. 2020), MTI-Net (Vandenhende et al. 2020), PSD (Zhou et al. 2020), and ATRC (Brüggemann et al. 2021). The performance comparison is shown in Table 4 and Table 5. It can be seen that our approach applied on the strong multi-task baseline achieves the highest performance on the different tasks and on both the NYUD-v2 and Pascal-Context datasets, demonstrating the superiority of the proposed approach. Some qualitative comparison with the baseline model is also shown in Fig. 6.

Conclusions

We presented the proposed cross-task contrastive learning model, a novel regularization for the multi-task dense prediction problem based on cross-task contrastive consistency on task-specific features, and also further introduced several important components designed for the proposed contrastive model, including an effective pixel sampling strategy, a generic positive/negative definition criterion for both continuous and discrete tasks, shared feature projection scheme, and contrastive task-pair selection to reduce overhead. Extensive experiments on NYUD-v2 and PASCAL-Context clearly verified the effectiveness of the proposed approach.

Acknowledgements

This research was partially supported by Early Career Scheme of the Research Grants Council (RGC) of the Hong Kong SAR under grant No. 26202321, and HKUST Startup Fund No. R9253.

References

Alonso, I.; Sabater, A.; Ferstl, D.; Montesano, L.; and Murillo, A. C. 2021. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *ICCV*.

Brüggemann, D.; Kanakis, M.; Obukhov, A.; Georgoulis, S.; and Van Gool, L. 2021. Exploring relational context for multi-task dense prediction. In *ICCV*.

Chaitanya, K.; Erdil, E.; Karani, N.; and Konukoglu, E. 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *NeruIPS*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *ICML*.

Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. E. 2020b. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020c. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Chen, Z.; Badrinarayanan, V.; Lee, C.-Y.; and Rabinovich, A. 2018. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *IJCV*, 88(2): 303–338.

Gao, Y.; Ma, J.; Zhao, M.; Liu, W.; and Yuille, A. L. 2019. Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. In *CVPR*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*.

Hu, H.; Cui, J.; and Wang, L. 2021. Region-Aware Contrastive Learning for Semantic Segmentation. In *ICCV*.

Kanakis, M.; Bruggemann, D.; Saha, S.; Georgoulis, S.; Obukhov, A.; and Gool, L. V. 2020. Reparameterizing convolutions for incremental multi-task learning without task interference. In *ECCV*.

Kendall, A.; Gal, Y.; and Cipolla, R. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*.

Kirillov, A.; Girshick, R.; He, K.; and Dollár, P. 2019. Panoptic feature pyramid networks. In *CVPR*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *CVPR*.

Liu, S.; Johns, E.; and Davison, A. J. 2019. End-to-end multi-task learning with attention. In *CVPR*.

Maninis, K.-K.; Radosavovic, I.; and Kokkinos, I. 2019. Attentive single-tasking of multiple tasks. In *CVPR*.

Neven, D.; De Brabandere, B.; Georgoulis, S.; Proesmans, M.; and Van Gool, L. 2017. Fast scene understanding for autonomous driving. *arXiv preprint arXiv:1708.02550*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*.

Sener, O.; and Koltun, V. 2018. Multi-task learning as multi-objective optimization. In *NeurIPS*.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *ECCV*.

Teichmann, M.; Weber, M.; Zoellner, M.; Cipolla, R.; and Urtasun, R. 2018. Multinet: Real-time joint semantic reasoning for autonomous driving. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, 1013–1020. IEEE.

Van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; and Van Gool, L. 2021. Unsupervised semantic segmentation by contrasting object mask proposals. In *ICCV*.

Vandenhende, S.; Georgoulis, S.; Gool, L. V.; and Gool, L. V. 2020. Mti-net: Multi-scale task interaction networks for multi-task learning. In *ECCV*.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *TPAMI*, 43(10): 3349–3364.

Xu, D.; Ouyang, W.; Wang, X.; and Sebe, N. 2018. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*.

Xu, D.; Ricci, E.; Ouyang, W.; Wang, X.; and Sebe, N. 2017. Multi-Scale Continuous CRFs as Sequential Deep Networks for Monocular Depth Estimation. In *CVPR*.

Ye, H.; and Xu, D. 2022. Inverted Pyramid Multi-task Transformer for Dense Scene Understanding. In *ECCV*.

Zhang, F.; Torr, P.; Ranftl, R.; and Richter, S. 2021. Looking Beyond Single Images for Contrastive Semantic Segmentation Learning. In *NeruIPS*.

Zhang, Z.; Cui, Z.; Xu, C.; Yan, Y.; Sebe, N.; and Yang, J. 2019. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*.

Zhao, X.; Vemulapalli, R.; Mansfield, P. A.; Gong, B.; Green, B.; Shapira, L.; and Wu, Y. 2021. Contrastive Learning for Label Efficient Semantic Segmentation. In *ICCV*.

Zhong, Y.; Yuan, B.; Wu, H.; Yuan, Z.; Peng, J.; and Wang, Y.-X. 2021. Pixel Contrastive-Consistent Semi-Supervised Semantic Segmentation. In *ICCV*.

Zhou, L.; Cui, Z.; Xu, C.; Zhang, Z.; Wang, C.; Zhang, T.; and Yang, J. 2020. Pattern-structure diffusion for multi-task learning. In *CVPR*.