Lifting the Veil of Frequency in Joint Segmentation and Depth Estimation

Tianhao Fu[†*]
tianhaofu1@gmail.com
Baidu

Yingying Li[†]
liyingying05@baidu.com
Baidu

Xiaoqing Ye[†]
yexiaoqing@baidu.com
Baidu

Xiao Tan
tanxchong@gmail.com
Baidu

Hao Sun
qianxun.sun@gmail.com
Baidu

Fumin Shen
fshen@uestc.edu.cn
University of Electronic Science and Technology of China

Errui Ding
dingerrui@baidu.com
Baidu

ABSTRACT

Joint learning of scene parsing and depth estimation remains a challenging task due to the rivalry between the two tasks. In this paper, we revisit the mutual enhancement for joint semantic segmentation and depth estimation. Inspired by the observation that the competition and cooperation could be reflected in the feature frequency components of different tasks, we propose a Frequency Aware Feature Enhancement (FAFE) network that can effectively enhance the reciprocal relationship whereas avoiding the competition. In FAFE, a frequency disentanglement module is proposed to fetch the favorable frequency component sets for each task and resolve the discordance between the two tasks. For task cooperation, we introduce a re-calibration unit to aggregate features of the two tasks, so as to complement task information with each other. Accordingly, the learning of each task can be boosted by the complementary task appropriately. Besides, a novel local-aware consistency loss function is proposed to impose on the predicted segmentation and depth so as to strengthen the cooperation. With the FAFE network and new local-aware consistency loss encapsulated into the multi-task learning network, the proposed approach achieves superior performance over previous state-of-the-art methods. Extensive experiments and ablation studies on multi-task datasets demonstrate the effectiveness of our proposed approach.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Multi-task learning**.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475277

KEYWORDS

Multi-task learning; Semantic segmentation; Depth estimation

ACM Reference Format:
Tianhao Fu[†*], Yingying Li[†], Xiaoqing Ye[†], Xiao Tan, Hao Sun, Fumin Shen, and Errui Ding. 2021. Lifting the Veil of Frequency in Joint Segmentation and Depth Estimation. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475277

1 INTRODUCTION

Thanks to the development of deep neural networks, the performances of vision tasks have been boosted to an unprecedented level. In the real world, we are able to solve multiple tasks at the same time and take advantage of the interrelation between variable subtasks. As a result, many efforts have been made on intelligent yet competent multi-task learning in recent years. In scene perception and understanding, depth estimation and semantic segmentation are two elementary problems as the former perceives 2.5D information for recovering the scene and the latter helps to conceive the scene. Although the deep-learning based methods have achieved great success in these two individual tasks [2, 4], the collaboration between the monocular depth estimation and semantic segmentation is overlooked. On the one hand, to make joint-optimization of the two tasks whereas restricting the computation head, a common practice is to share the same encoder and apply a multi-head architecture for regression. On the other hand, due to internal competition and trade-offs, sub-optimal performance for each task is usually observed.

Conventional deep multi-task learning approaches of monocular depth estimation and semantic segmentation aim at sharing representations between them [45]. But due to the competition between the two tasks, joint learning of segmentation and depth estimation remains a challenging task. Some works have been proposed to take

[†]Equal Contribution.
[*]Corresponding author: Tianhao Fu.

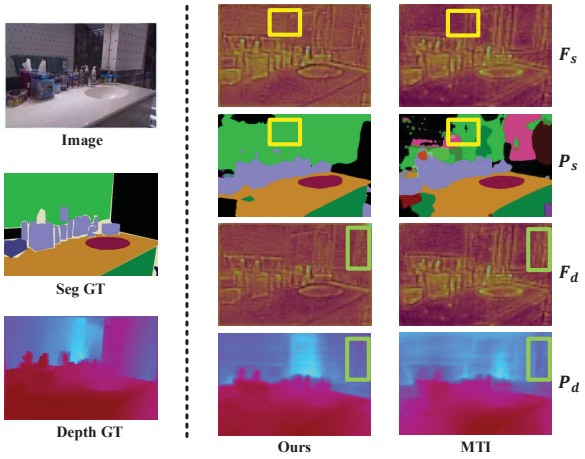

Figure 1: Features and results of MTI-Net [37] and ours. *F* **represents the feature map and** *P* **represents prediction of each task.***GT* **represents ground truth of each task.** *MTI* **represents our baseline MTI-Net.** *Ours* **represents our FAFE-Net.** *s* **and** *d* **denote segmentation and depth estimation respectively. For the features in an object, the yellow box indicates that our segmentation feature within an object has more low-frequency components, resulting in better segmentation results. Similarly, the green box shows that the depth feature in an object is much smoother in our results than MTI-Net.**

frequency domain knowledge into consideration in computer vision tasks, such as those for designing attention modules [34], for solving super-resolution problems [41], and for semantic segmentation tasks [19] and etc. However, how to exploit the frequency domain information to leverage the performance of multi-task learning, remains an open question.

In this paper, we manage to solve the problem of collaboration between tasks in a perspective of the frequency domain. Firstly, we reveal that there is not only correlation but also competition between the two tasks. On the one hand, promoting their correlation can get mutual enhancement, on the other hand, we need to avoid the competition between them to get better task-dependent representations. Secondly, the two tasks should focus on the different frequency bands in the domain. As shown in Figure 1, the features are consistent with the prediction results. For example, to get a better segmentation performance, the segmentation task should pay more attention to high-frequency features at the edge of an object and more low-frequency features within an object. For the depth estimation task, the low-frequency feature can smooth out the effect, while the high-frequency feature can deal with situations where there is a depth jump.

Besides the network architecture, we also consider the relationship of segmentation and depth estimation from the perspective of loss function. Intuitively, for the same object, the corresponding pixels on the segmentation map should have the same class label, whereas the values on the depth map should also be similar in a local region.

Based on this, we propose a local-aware consistency loss, such that for a certain area of the same object, the variance distribution of segmentation and depth predictions within a local area should be consistent, which could further improve the multi-task learning performance.

To summarize, the contribution of this paper is:

- For multi-task learning, we propose a Frequency Aware Feature Enhancement (FAFE) network that can effectively enhance the reciprocal relationship whereas avoiding the competition in the frequency domain. We introduce the FAFE network building upon two main modules: frequency disentanglement module and a feature re-calibration unit.
- We propose a new loss called local-aware consistency loss based on the analysis of that semantic segmentation task results should be consistent with depth estimation results within a local region of the same object.
- Extensive experiments on the challenging NYUD-v2 and Cityscapes datasets demonstrate the effectiveness of the proposed approach. Our approach achieves state-of-the-art results on NYUD-v2 and Cityscapes on jointly optimizing both the depth estimation and the segmentation tasks.

2 RELATED WORKS

Semantic segmentation and Depth estimation. Semantic segmentation is a high-level vision task that aims to facilitate per-pixel label classification. Recent methods designed for semantic segmentation are mainly based on deep neural networks. Long *et al.* [27] is the pioneering work that leverages a fully convolutional network (FCN) to achieve remarkable segmentation performance. Subsequently, many segmentation networks have been proposed such as DeepLab series [3–5], PSPNet [46], and so on, which take advantage of atrous convolutions and pyramid module to effectively increase the receptive field and fuse convolutional features from multiple scales.

Many efforts have been devoted to monocular depth estimation task. Previous methods are generally based on hand-crafted features and graphical models like Markov Random Field (MRF) [31, 32]. Recently some works adapt image classification networks into fully convolutional forms to predict depth on different sizes of inputs [8, 23]. Adabins [2] proposed to divide the depth range into different bins, the final depth values are estimated as linear combinations of the bin centers. These works only focused on an individual task without jointly optimizing the depth estimation and scene parsing together.

Multiple Task Learning (MTL). Many computer vision problems are multi-modal, for example, it is expected to segment the lane markings, detect vehicles, estimate depth in autonomous driving. The multi-modal requirements have motivated researchers to develop Multiple Task Learning (MTL), and deep MTL has been widely used so that a deep learning model can infer all desired task outputs [6, 10, 16]. There are a lot of successful multi-task pairs that have yielded fruitful results such as segmentation and depth estimation [37, 42], classification and detection [11, 29], detection and segmentation [7, 13], and so on.

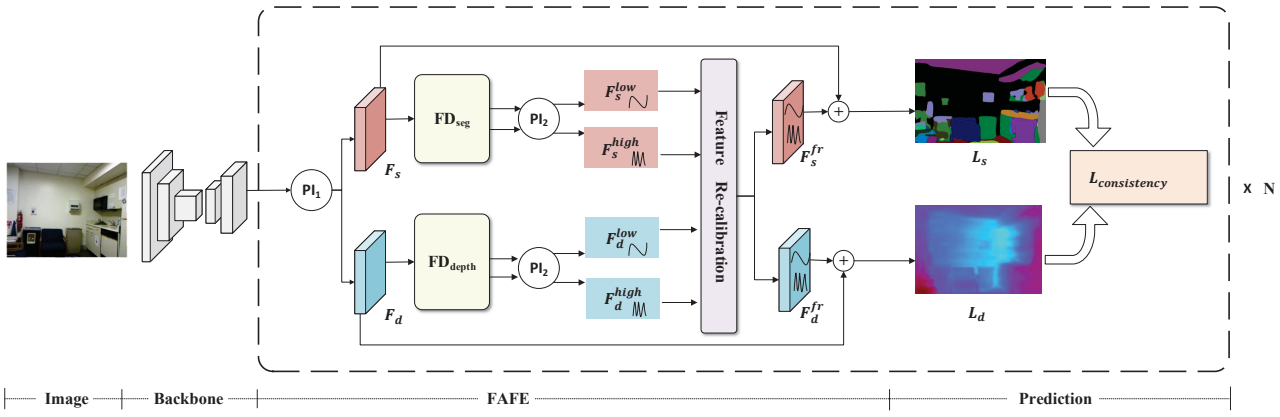

Figure 2: Architecture overview. Our Frequency Aware Feature Enhancement (FAFE) network contains two main modules: a Frequency Disentanglement (FD) module to extract the high-frequency and low-frequency features of each task respectively and a Feature Re-calibration (FR) unit to make full use of the multi-task learning advantages. *PI* means the pre-interaction module and the subscripts 1 and 2 means single-input and dual-input. *FD* denotes frequency disentanglement module. *L* denotes the loss, $L_{consistency}$ is our local-aware consistency loss. Subscripts *s* and *d* represent segmentation and depth estimation task respectively. Superscripts *low* and *high* represent low-frequency and high-frequency, respectively. Our approache could also be easily extended to *N* predictions in order to add progressive supervision. This can be done simply by embedding our FAFE network in front of each prediction.

The main topic in MTL research is to improve generalization of the original single task by sharing representations between related tasks [36]. Some works aim to improve different task features diffusion to strengthen task performance. MTAN [26] proposed a soft-attention module for each task so that each task has its task-specific feature. PSD [47] propose a pattern-structure diffusion framework to mine and propagate task-specific and task-across pattern structures in the task-level space for Multi-task Learning. MTI-Net [37] proposed to create interaction in different scales between affinity tasks to make the architecture learn more representative features.

However, the aforementioned works do not analyze the rivalry between the tasks. CSTRACK [20] analyzed the joint learning detection and ReID task, which reveals that the competition of them inevitably hurts the learning of task-dependent representations, they come up with a decouple module when split task-specific feature from source common feature. It not only effectively mitigate the competence, but also improve the collaborative learning capability between different tasks. Meanwhile, decouple module is also to be used in some single tasks to further improve feature representation [25, 40, 44].

A series of classic operations of deep convolution neural networks, such as feature aggregation and some attention networks, can be treated as implicit modeling of the frequency domain. Early works are combining traditional frequency decomposition methods like Wavelet Transform (WT) [41] and Discrete Cosine Transform (DCT) [39] to explicitly model features in the frequency domain. Recently, FCANet [34] makes a detailed derivation of DCT. In the segmentation task, li *et al.* [19] uses decoupled supervision to model the object body and edge, which correspond to the high and low-frequency of the image. In our join segmentation and depth estimation task, we believe each task needs to explicitly extract different frequency band information, and the interaction between low-frequency and high-frequency of these two tasks should be considered to disentanglement into different feature distributions for each task.

3 METHOD

In this section, we describe the proposed Frequency Aware Feature Enhancement (FAFE) network for simultaneous depth estimation and scene parsing. We first present an overview of our network architecture in section 3.1, and then introduce the details of the FAFE network in section 3.2. Finally, the local-aware consistency loss function is illustrated in section 3.3.

3.1 Network Architecture

The joint segmentation and depth estimation network we used could be based on any backbone such as HRNet [35]. Figure 2 is the network architecture overview, which shows that our Frequency Aware Feature Enhancement (FAFE) network is inserted before the multi-task predictions. Similarly, in recent multi-task networks such as PAD-Net [47] which have two stage predictions, our FAFE network is inserted before intermediate predictions and final predictions separately. And MTI-Net [37] has a backbone that extracts multi-scale features and intermediate predictions are made at each scale, so our FAFE network is inserted before the intermediate predictions at each scale.

3.2 Frequency Aware Feature Enhancement (FAFE) Network

Our Frequency Aware Feature Enhancement (FAFE) network is designed following two guidelines. First, it should make different branches learn task-specific features in the frequency domain since distinct tasks typically encourage learning different types of features corresponding to a particular frequency. Second, it is designed for making one task could get information from the other task which the former task needed but don't include. To meet the first requirement, the feature map generated by the backbone is firstly transferred into different feature maps for different tasks by an interaction module inspired from [20], and then a frequency disentanglement module is followed to learn the high-frequency and low-frequency features of each task respectively. This disentanglement module is capable to generate task-specific features. Our frequency disentanglement module is shown in Figure 3. For the second target, we use an attention-based neural network to fuse the corresponding frequency features of the other task, as shown in Figure 5. To sum up, Our FAFE network, which consists of Pre-Interaction (PI), Frequency Disentanglement (FD) and Feature Re-calibration (FR) modules, solves the competition and cooperation trade-off in multi-task learning effectively.

Pre-interaction (PI) Module. Our pre-interaction (PI) module is derived from the cross correlation module of CSTrack [20], which consists of a intra-attention part and a inter-attention part. The original cross correlation module has one input and two outputs. We expand it so that the module can accept two inputs. Specifically, the two inputs are respectively passed through two different convolution layers, followed by an avg-pooling operation to generate a fixed-size feature map, the subsequent attention computation is the same as the original cross correlation module. In FAFE, we use single-input-PI to complete the initial interaction, and the dual-input-PI is used after the frequency disentanglement module to get better feature representation by fusing different frequency feature maps. Take dual-input-PI for example, the implementation steps are as follows:

$$\begin{aligned} V_1, V_2 &= func_1(I_1), func_2(I_2) \\ M_1 &= \alpha_1 \cdot (V_1^\top \cdot V_1) + (1-\alpha_1) \cdot (V_1^\top \cdot V_2) \\ M_2 &= \alpha_2 \cdot (V_2^\top \cdot V_2) + (1-\alpha_2) \cdot (V_2^\top \cdot V_1) \\ O_1, O_2 &= V_1 \cdot M_1 + V_1, V_2 \cdot M_2 + V_2 \end{aligned} \tag{1}$$

where I_1 and I_2 stand for the dual inputs, $func_1$ and $func_2$ contain a convolution layer followed by an average -pooling operation and a reshape operation respectively. For one-input-PI, I_1 is equal to I_2 and $func_1$ is equal to $func_2$. M_1 and M_2 represent the attention weight matrix of V_1 and V_2, α_1 and α_2 are the balanced weights of intra-attention and inter-attention. O_1 and O_2 are the final outputs after interaction.

Frequency Disentanglement (FD). As demonstrated in [34], it is beneficial to convert features in the spatial domain to the frequency domain. First of all, since the original signal can be recovered perfectly by the transferred signal using inverse transformation, it is safe to utilize features in the frequency domain to do analysis without worrying about information loss. Moreover, the strength of the frequency feature lies in the fact that its energy arrangement is more compact, and it is well known that the natural

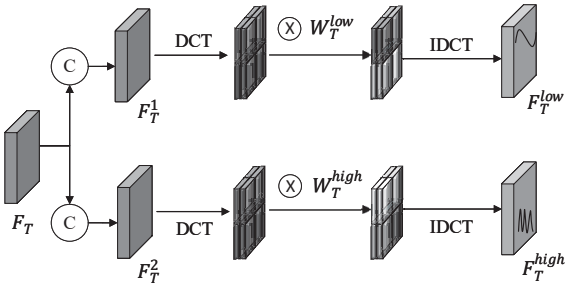

Figure 3: Frequency disentanglement module. In FD module, we use discrete cosine transform (DCT) to get each frequency component, and then two weight matrices are learned for each task to get low and high-frequency features. c means convolution operation. T represent semantic segmentation task or depth estimation task. *low* and *high* represents the low-frequency and high-frequency respectively.

image is frequency band limited and hence we can expect feature manipulation in the frequency domain is more robust against its counterparts in the spatial domain, and this might be the reason that recent works of exploiting frequency transformation in neural network achieve great success [19, 34]. Although simple, the frequency band is divided more finely so that different frequency bands can be chooses.

For each task, we generate two feature maps using two convolution layers, representing low-frequency branch and high-frequency branch respectively. And then we use a two-dimensional (2D) DCT to get 2D DCT frequency spectrum, which can be written as:

$$\begin{aligned} f_{h,w}^{2d} = \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x_{i,j}^{2d} cos(\frac{\pi h(i+0.5)}{H}) cos(\frac{\pi w(j+0.5)}{W}), \\ s.t. h \in (0,1,...,H-1), w \in (0,1,...,W-1) \end{aligned} \tag{2}$$

$x_{i,j}^{2d}$ is the pixel value at location (i, j). $cos()$ represents the cosine function. H and W are the height and the width of the inputs respectively. The value at each position of the frequency spectrum $f_{h,w}^{2d}$ represents corresponding frequency component. For low-frequency/high-frequency branch, we use an adaptive weight matrix $W_{h,w}^{2d}$ to learn the proportion of the frequency components. The details of extraction procedure is presented in Figure 4. After that, we get low-frequency/high-frequency feature map through inverse DCT (IDCT).

The whole process of FD module can be expressed as follows:

$$\begin{aligned} F_T^{low} &= IDCT(DCT(conv_1(F_T)) \cdot W_T^{low})) \\ F_T^{high} &= IDCT(DCT(conv_2(F_T)) \cdot W_T^{high})) \end{aligned} \tag{3}$$

where the subscript T can represent semantic segmentation task or depth estimation task, F_T represents the input feature map while *conv* represents convolution operation. *DCT* and *IDCT* represents Discrete Cosine Transform and Inverse Discrete Cosine Transform

respectively. W_T^{high} and W_T^{low} are the adaptive weight matrices of low-frequency and high-frequency. F_T^{low} and F_T^{high} represent the output feature maps of low-frequency and high-frequency.

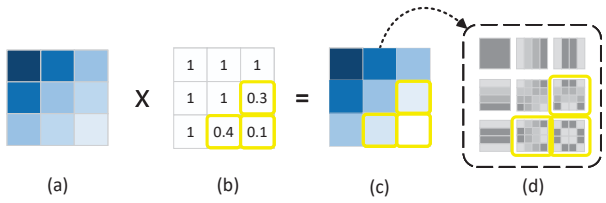

Figure 4: Illustration of using adaptive weight matrix to get low/high frequency components. (a) Original 2D DCT frequency spectrum. (b) Adaptive weight matrix. (c) Weighted 2D DCT frequency spectrum. (d) Visualization of frequency.

As described above, we get the initial low-frequency feature map and high-frequency feature map. And then these two feature maps are processed by the dual-input-PI module to get better low and high frequency feature representation. Each task has learned its own low-frequency and high-frequency features after completing the frequency disentanglement module, which alleviate rival effectively.

Feature Re-calibration (FR). It is worth noticing that the features extracted by the frequency disentanglement module alone considerate only task-specific features without taking collaboration between the segmentation and the depth estimation tasks into consideration. We argue that extracting features by modeling the collaboration between tasks will be beneficial. For example, some edges detected in the high-frequency feature maps of segmentation are also helpful for the depth estimation task to decide where are the depth boundaries. To achieve this goal, the low-frequency features of the segmentation task are sent to the SE attention module [14] to generate the re-calibration features, which is added back to the low-frequency features of the depth estimation task, so that the depth estimation task can learn some useful low-frequency features from segmentation features. We implement the feature re-calibration of other frequency features in the same way which could further improves both the segmentation and depth estimation performance. The FR unit is shown in detail in Figure 5.

3.3 Local-aware Consistency Loss

Following the loss function settings of MTI-Net [37], we use the L_1 loss for depth estimation and the cross-entropy loss for semantic segmentation. Besides these, we propose a local-aware consistency loss in order to dig the similarity features of segmentation results and depth estimation results. For the segmentation task, the prediction results of the same object are expected to be consistent. Meanwhile, the depth estimation results within an object should be smooth.Based on this intuition, the local-aware consistency loss is proposed to learn the smooth similarities between them. For each pixel in segmentation prediction P_s, we calculate the mean of the 3x3 neighborhood centered at this point and generate a mean map MP_s. Then we use the difference P_s and MP_s to get consistency map of segmentation C_s. In C_s, the value of each pixel represents the segmentation consistency of the 3x3 neighborhood. If the 3x3 neighborhood belongs to the same object, then the corresponding value in C_s should be very small.

Similar to segmentation, we can get depth consistency map C_d, which represents the smoothness of depth prediction. It is worth noting that if the pixel value of C_s is too large, this location is considered to be the edge of an object, so we don't take the consistency of this location into consideration.

We propose that the smoothness of segmentation and depth should be consistent, so our local-aware consistency loss is calculated by L_1 loss between the C_s and C_d. The consistency loss and the whole loss are formulated as follows:

$$\begin{aligned} MP_T^i &= \frac{1}{N} \sum_{j \in \mathcal{N}(P_T^i)} P_T^j \\ C_T^i &= (P_T^i - MP_T^i)^2 \\ L_c &= L_1(C_s, C_d) \\ L &= \lambda_1 \cdot L_s + \lambda_2 \cdot L_d + \lambda_3 \cdot L_c \end{aligned} \tag{4}$$

in which i denotes a particular pixel in a map. T means the corresponding task like segmentation or depth estimation. $\mathcal{N}(P_T^i)$ represents the neighbor set of P_T^i. L_s is cross-entropy loss for semantic segmentation and L_d is the L_1 loss for depth estimation. L_c is the local-aware consistency loss. $\lambda_1, \lambda_2, \lambda_3$ is the weight of each loss and we set $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 0.1$.

4 EXPERIMENT

4.1 Experimental setup

Datasets. We perform our experimental evaluation on NYUD-v2 [33] and Cityscapes [21]. The NYUD-v2 dataset is a popular indoor-scene image dataset, which has been widely used for depth estimation [9] and semantic segmentation [12]. Following the previous work MTI-Net [37], we use 795 images for training and 654 images to test the final performance. In addition, we also adopt the data augmentation method of MTI-Net [37] to get the augmented training data. The Cityscapes dataset consists of outdoor scene images with overall 19 semantic classes annotated for semantic segmentation. Besides, Cityscapes also provides pre-computed disparity maps which can be regarded as inverse depth labels. Similar to MTAN [26], we use 7 categories CityScapes dataset and resize the images to 128 × 256 to speed up the forward stage. The training data are augmented following the previous work [37].

Implementation details. We build our framework based on HRNet [35] backbone, with ImageNet-pretrained HRNet-18 for ablation studies and HRNet-48 for the final results. The network is trained for joint depth estimation and segmentation tasks in an end-to-end manner. We use the Adam optimizer with initial learning rate 1e-4, and batch size is set to 6. Totally 80 epochs are used for NYUD-v2, and 200 epochs for Cityscapes. Note that when computing local-aware consistency loss, we discard large values in C_s and C_d in Equation 4 since these areas represent low consistency between depth and segmentation and should be ignored. We implement it by using a mask to neglect large value when computing the loss. In particular, we set the threshold to 1.

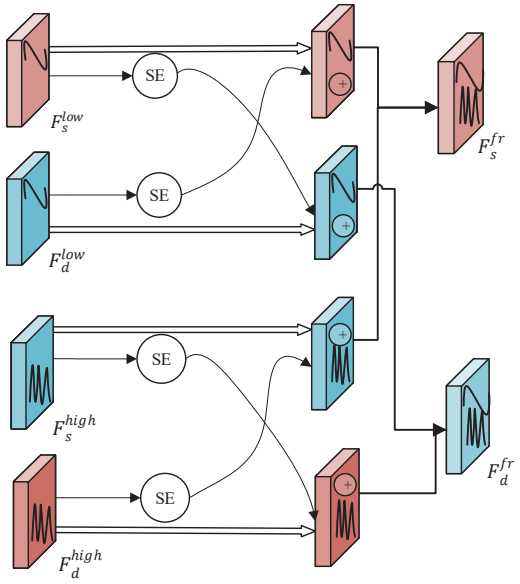

Figure 5: Feature re-calibration module. *SE* **is the SE attention [14]. The plus sign denotes adding the re-calibration feature from the other task. After feature re-calibration, the feature map of segmentation task F_s^{fr} integrates the frequency information which is extracted by depth estimation task while F_d^{fr} is also learned from the segmentation task branch.**

Evaluation Metrics. For the evaluation of semantic segmentation, we take the same metrics as [37, 42]: mean intersection over union (mIoU), mean accuracy (mAcc) and pixel accuracy (PixAcc). For the evaluation of depth estimation, we take the same metrics as [1, 17, 26] : root mean squared error (rmse), average relative error (rel), mean absolute error (mae), threshold accuracy (δ_i) where threshold = $1.25, 1.25^2, 1.25^3$. Specifically, on NYUD-v2 dataset, we follow MTI-Net [37], which mainly focuses on the root mean squared error (rmse). For Cityscapes dataset, we mainly focus on average relative error (rel) by following MTAN [26].

Baselines. On both NYUD-v2 and Cityscapes datasets, we compare our FAFE network with several state-of-the-art multi-task learning methods, such as PAD-Net [42] and MTI-Net [37]. Noted that these two methods both have two intermediate auxiliary tasks, which is surface normal estimation and contour prediction. For fair comparison, we use the open-source code to reproduce their works on the joint two tasks involving segmentation and depth estimation, instead of four sub-tasks. Besides, the multi-task network that learned more than two tasks like PAP-Net [47] without open-source code is not included in our comparison.

Table 1: Ablation studies on NYUD-v2 dataset. + in first column denote the basis network plus our method component while + in second column and third column denote the absolute performance improvement. ↑ represents big number is better and ↓ represents small number is better. The contents in parentheses in the first row indicate the metrics we used. Same rules apply to the rest table.

Method	Seg (mIoU) ↑	Depth (rmse) ↓
Baseline (PAD)	35.53	0.627
+PI	37.69 (+2.16)	0.570 (+0.057)
+PI+FD	38.22 (+2.69)	0.560 (+0.067)
+FAFE (PI+FD+FR)	38.31 (+2.78)	0.558 (+0.069)
+FAFE+L_c	**39.24 (+3.71)**	**0.554 (+0.073)**
Baseline (MTI)	37.19	0.549
+PI	38.46 (+1.27)	0.552 (-0.003)
+PI+FD	39.04 (+1.85)	0.534 (+0.015)
+FAFE (PI+FD+FR)	39.59 (+2.40)	0.531 (+0.018)
+FAFE+L_c	**39.91 (+2.72)**	**0.529 (+0.020)**

4.2 Ablation studies

In Table 1 and 2 we reveal the results of our ablation studies on NYUD-v2 and Cityscapes datasets, respectively. We conduct experiments on two different network architectures, i.e., PAD-Net [42] and MTI-Net [37], to verify the generalization capability and contribution of FAFE components.

We focus on the NYUD-v2 dataset first (see Table 1), HRNet-18 backbone is used for PAD-Net [42] and MTI-Net [37]. The PAD-Net baseline has lower performance (mIoU is 35.53 and rmse is 0.627) than the MTI-Net baseline (mIoU is 37.19 and rmse is 0.549). However, PAD-Net's segmentation accuracy even outperforms the original MTI-Net after adopting our FAFE module and our proposed consistency loss function (L_c). As can be seen from the yellow-colored row, the mIoU of segmentation is 39.24 and the rmse of depth estimation is 0.554. As we mentioned above, frequency disentanglement(FD) module could decouple entangled features in frequency domain, feature Re-calibration module could make different task features learn from each other to get more useful information. Local-consistency loss could make the segmentation result more better and depth result more smoother. We also have proved the effectiveness of each component whether on PAD-Net baseline or MTI-Net baseline,which shown in Table 1. Take the MTI-Net as an example, the effect has been improved for both segmentation (+1.85) and depth estimation (+0.015) after adding frequency disentanglement (FD) module. When including the feature re-calibration (FR) unit, another significant boost in performance is achieved (seg: +2.40, depth: +0.018). Finally, using the auxiliary consistency loss (L_c) can further help to improve the performance of our predictions (seg: +2.72, depth: +0.020).

Table 2 shows the ablation on Cityscapes dataset. Similar to NYUD-v2 dataset, the baseline networks are PAD-Net and MTI-Net and we adopted the same the backbone we used is HRNet-18. Being consistent with the performance on NYUD-v2, Table 2 verified the effectiveness of our frequency aware feature enhancement (FAFE) network and the new consistency loss (L_c). For PAD-Net baseline,

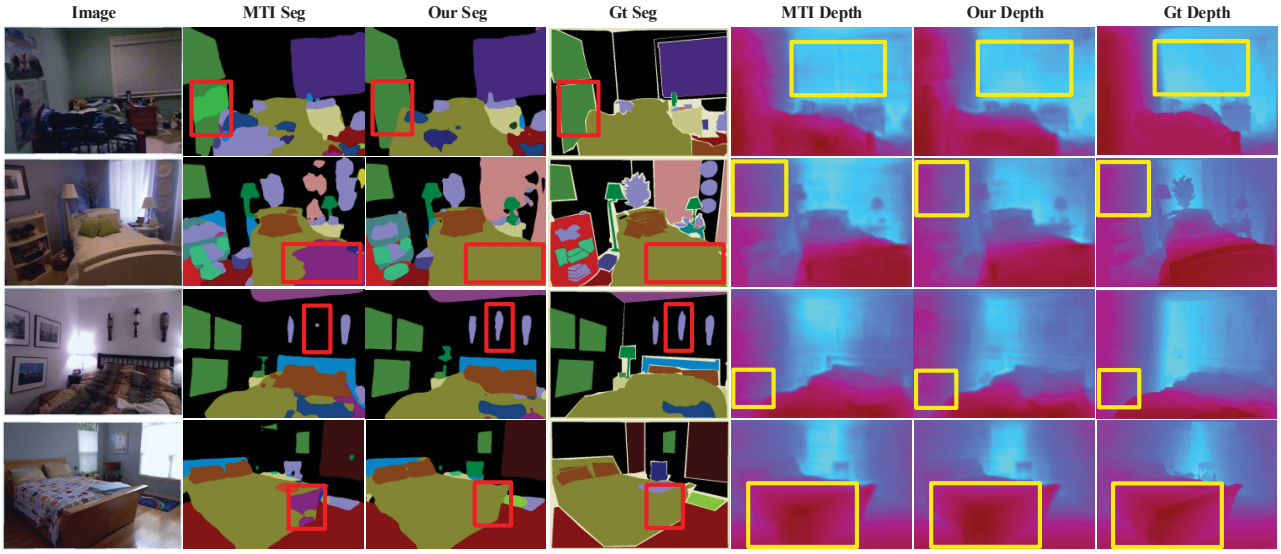

Figure 6: Qualitative results of our FAFE network on NYUD-v2 dataset. *MTI Seg* **and** *MTI Depth* **represents MTI-Net segmentation result and depth result respectively.** *Our Seg* **and** *Our Depth* **represents FAFE-Net segmentation result and depth result respectively.**

Table 2: Ablation studies on Cityscapes dataset.

Method	Seg (mIoU)↑	Depth (rel)↓
Baseline (PAD)	77.20	56.761
+PI	77.24 (+0.04).	43.409 (+13.352)
+PI+FD	77.40 (+0.20)	35.443 (+21.318)
+FAFE (PI+FD+FR)	77.45 (+0.25)	34.872 (+21.889)
+FAFE+L_c	**77.61 (+0.41)**	**31.011 (+25.750)**
Baseline (MTI)	76.02	52.902
+PI	76.04 (+0.02)	49.856 (+3.042)
+PI+FD	76.17 (+0.15)	47.393 (+5.509)
+FAFE (PI+FD+FR)	76.22 (+0.20)	45.336 (+7.566)
+FAFE+L_c	**76.26 (+0.24)**	**41.932 (+10.970)**

we observe 0.41 improvement on segmentation and 25.750 improvement on depth estimation. For MTI-Net baseline, our results show that the segmentation improves 0.24 in mIoU criteria and the depth estimation observes 10.970 gain in rel criteria. The two indoor and outdoor datasets validate the generalization ability of the proposed approach.

Influence of prediction heads.

From Figure 2, we could observe that our Frequency Aware Feature Enhancement (FAFE) network is inserted before the multi-task predictions, which means that our whole network architecture could have multiple prediction heads. In order to verify the effectiveness of our FAFE network on different architectures, we conduct experiments to analyse the results of the our FAFE network with different numbers of prediction heads. The models are based on HRNet-18 backbone and trained on NYUD-v2 dataset, and the connection module between multiple prediction heads is borrowed from PAD-Net [42]. As illustrated in Table 3, we can see that the performances are generally improved after using our FAFE network from one prediction head to three prediction heads. When the number of prediction heads is 2, it has the highest performance with our FAFE network. One of the potential reasons is that the two prediction heads in our FAFE network is enough to help the network to learn the task-specific features.

Table 3: Ablating the prediction heads on NYUD-v2.

Method	seg (mIOU)↑	depth (rmse) ↓
1-predict	35.40	0.580
1-predict+FAFE+L_c	36.49 (+1.09)	0.563 (+0.017)
2-predict	35.53	0.627
2-predict+FAFE+L_c	39.24 (+3.71)	0.554 (+0.073)
3-predict	35.88	0.600
3-predict+FAFE+L_c	38.12 (+2.24)	0.548 (+0.052)

Influence of the parameters. In Table 4 we visualize the results of our ablation studies on NYUD-v2 using HRNet-18 backbone to verify whether it is our FAFE network or the extra convolution parameters that contributes to the multi-task improvements. We removed some innovative operations such as 2D DCT for frequency analysis, matrix multiplication for attention work. At the same time, we keep the convolution operations in order to keep the same parameters. As shown in Table 4, compared to the network with the same parameters, our FAFE shows a significant positive effect on both segmentation and depth estimation results.

Table 4: Influence of the parameters on NYUD-v2.

Method	Params (M)	Seg (mIoU)↑	Depth (rmse)↓
PAD+conv	17.67	36.27	0.543
PAD+FAFE	17.67 (+0)	39.59 (+3.32)	0.531 (+0.012)

Table 5: Comparison with the state-of-the-art on NYUD-v2.

(a) Results on depth estimation.

Method	rmse↓	rel↓	mae↓	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
HCRF [18]	0.821	0.232	-	0.621	0.886	0.968
DCNF [24]	0.824	0.230	-	0.614	0.883	0.971
Wang [38]	0.745	0.220	-	0.605	0.890	0.970
NR forest [30]	0.774	0.187	-	-	-	-
Xu [43]	0.593	0.125	-	0.806	0.952	0.986
PAD-Net	0.485	0.139	0.361	0.814	0.962	0.992
MTI-Net	0.473	0.140	0.359	0.824	0.964	0.992
Ours	**0.466**	0.135	**0.344**	**0.834**	**0.967**	**0.992**

(b) Results on semantic segmentation.

Method	PixAcc↑	mAcc↑	mIoU↑
FCN [27]	60.0	49.2	29.2
Context [22]	70.0	53.6	40.6
Eigen [8]	65.6	45.1	34.1
B-SegNet [15]	68.0	45.8	32.4
RefineNet-101 [21]	72.8	57.8	44.9
PAD-Net	72.92	57.61	44.79
MTI-Net	72.71	58.31	47.05
Ours	**74.62**	**61.33**	**48.40**

4.3 Comparison with the State-of-the-Arts

In this section, we compare our proposed method with various state-of-the-art methods for Multi-task learning.

Comparison on NYUD-v2. Table 5 shows the comparison with the state-of-the-art approaches on NYUD-v2. We use the multi-scale HRNet-48 backbone and the PAD-Net and MTI-Net are reproduced by removing two intermediate auxiliary tasks for fair comparison. Since MTI-Net outperforms PAD-Net, we complete our model based on MTI-Net. As is shown in Table 5, the results compared to MTI-Net well demonstrates that our proposed approach can boost the performance of previous works and achieves new SoTA results on both depth estimation and semantic segmentation tasks. Figure 6 shows qualitative examples of the depth estimation and segmentation. We can observe that the regions within the red boxes indicate our segmentation results are more consistent within the object than the naive MTI without FAFE, and the yellow boxes indicate that our depth results are smoother than MTI without FAFE.

Comparison on Cityscapes. Similar to NYUD-v2, the results of PAD-Net and MTI-Net are reproduced on two tasks of segmentation and depth estimation. In contrast to NYUD-v2, the effect of PAD-Net is better than MTI-Net on Cityscapes dataset, so our model is based on the PAD-Net. In semantic segmentation task, our FAFE network achieves the best performance in all of the measure metrics. However, since our baseline has already achieved a high performance, the improvement of segmentation results is not particularly obvious. In depth estimation task, our model also achieves the best results especially on the average relative error (rel) index. The qualitative examples of the depth estimation and segmentation on Cityscapes are shown in Figure 7.

Table 6: Comparison with the state-of-the-art on Cityscapes.

(a) Results on depth estimation.

Method	rmse↓	rel↓	mae↓
Cross-Stitch[28]	-	34.49	0.0154
MTAN[26]	-	33.63	0.0144
MTI-Net	0.023	44.415	0.012
PAD-Net	0.023	29.650	0.011
Ours	**0.023**	**28.833**	**0.011**

(b) Results on semantic segmentation.

Method	PixAcc↑	mAcc↑	mIoU↑
Cross-Stitch[28]	90.33	-	50.08
MTAN[26]	91.11	-	53.04
MTI-Net	94.42	85.35	78.37
PAD-Net	94.47	85.72	78.63
Ours	**94.49**	**85.87**	**78.75**

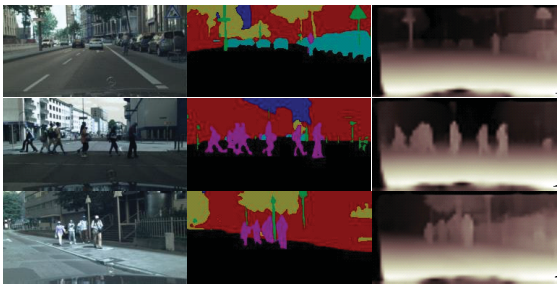

Figure 7: Qualitative results of our FAFE network on Cityscapes dataset.

5 CONCLUSION

In this paper, we proposed the Frequency Aware Feature Enhancement (FAFE) network and a local-aware consistency loss for joint segmentation and depth estimation. The FAFE network architecture consists of a frequency disentanglement module and a feature re-calibration unit, which can solve competition between segmentation and depth estimation while enhance the collaboration between the two tasks in an end-to-end manner. Besides, the local-aware consistency loss could improve these two tasks performance further through strengthen collaboration between tasks. Experiments on the NYUD-v2 and CityScapes datasets show that our method is competitive with other methods. In the future, we may generalize and improve the efficiency of our approach on other different tasks.

REFERENCES

[1] Ibraheem Alhashim and Peter Wonka. 2018. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941* (2018).

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. 2020. AdaBins: Depth Estimation using Adaptive Bins. *arXiv preprint arXiv:2011.14141* (2020).

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062* (2014).

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).

[6] Carl Doersch and Andrew Zisserman. 2017. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 2051–2060.

[7] Nikita Dvornik, Konstantin Shmelkov, Julien Mairal, and Cordelia Schmid. 2017. Blitznet: A real-time deep network for scene understanding. In *Proceedings of the IEEE international conference on computer vision*. 4154–4162.

[8] David Eigen and Rob Fergus. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*. 2650–2658.

[9] David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283* (2014).

[10] Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 109–117.

[11] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

[12] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. 2014. Learning rich features from RGB-D images for object detection and segmentation. In *European conference on computer vision*. Springer, 345–360.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[14] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[15] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. 2015. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680* (2015).

[16] Abhishek Kumar and Hal Daume III. 2012. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417* (2012).

[17] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. 2019. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326* (2019).

[18] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. 2015. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1119–1127.

[19] Xiangtai Li, Xia Li, Li Zhang, Guangliang Cheng, Jianping Shi, Zhouchen Lin, Shaohua Tan, and Yunhai Tong. 2020. Improving semantic segmentation via decoupled body and edge supervision. *arXiv preprint arXiv:2007.10035* (2020).

[20] Chao Liang, Zhipeng Zhang, Yi Lu, Xue Zhou, Bing Li, Xiyong Ye, and Jianxiao Zou. 2020. Rethinking the competition between detection and ReID in Multi-Object Tracking. *arXiv preprint arXiv:2010.12138* (2020).

[21] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1925–1934.

[22] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. 2016. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3194–3203.

[23] Fayao Liu, Chunhua Shen, and Guosheng Lin. 2015. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5162–5170.

[24] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. 2015. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence* 38, 10 (2015), 2024–2039.

[25] Jianbo Liu, Yongcheng Liu, Ying Wang, Véronique Prinet, Shiming Xiang, and Chunhong Pan. 2020. Decoupled representation learning for skeleton-based gesture recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5751–5760.

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

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[28] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. 2016. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3994–4003.

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* (2015).

[30] Anirban Roy and Sinisa Todorovic. 2016. Monocular depth estimation using neural regression forest. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5506–5514.

[31] Ashutosh Saxena, Sung H Chung, Andrew Y Ng, et al. 2005. Learning depth from single monocular images. In *NIPS*, Vol. 18. 1–8.

[32] Ashutosh Saxena, Min Sun, and Andrew Y Ng. 2008. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence* 31, 5 (2008), 824–840.

[33] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*. Springer, 746–760.

[34] Vivek Kumar Singh, Mohamed Abdel-Nasser, Hatem A Rashwan, Farhan Akram, Nidhi Pandey, Alain Lalande, Benoit Presles, Santiago Romani, and Domenec Puig. 2019. FCA-net: Adversarial learning for skin lesion segmentation based on multi-scale features and factorized channel attention. *IEEE Access* 7 (2019), 130552–130565.

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

[36] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. 2021. Multi-Task Learning for Dense Prediction Tasks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[37] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. 2020. Mti-net: Multi-scale task interaction networks for multi-task learning. In *European Conference on Computer Vision*. Springer, 527–543.

[38] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. 2015. Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2800–2809.

[39] Zhangyang Wang, Ding Liu, Shiyu Chang, Qing Ling, Yingzhen Yang, and Thomas S Huang. 2016. D3: Deep dual-domain based fast restoration of JPEG-compressed images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2764–2772.

[40] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. 2020. Label decoupling framework for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13025–13034.

[41] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. 2020. Invertible image rescaling. In *European Conference on Computer Vision*. Springer, 126–144.

[42] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. 2018. 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*. 675–684.

[43] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. 2018. Structured attention guided convolutional neural fields for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3917–3925.

[44] Hao Zhang, Mengmeng Wang, Yong Liu, and Yi Yuan. 2020. FDN: Feature Decoupling Network for Head Pose Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12789–12796.

[45] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. 2018. Joint task-recursive learning for semantic segmentation and depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 235–251.

[46] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2881–2890.

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