Efficient Multitask Dense Predictor via Binarization

Yuzhang Shang\textsuperscript{1}, Dan Xu\textsuperscript{2}, Gaowen Liu\textsuperscript{3}, Ramana Rao Kompella\textsuperscript{3}, Yan Yan\textsuperscript{1,†}
\textsuperscript{1}Illinois Institute of Technology, \textsuperscript{2}HKUST, \textsuperscript{3}Cisco Research
yzshawn@outlook.com, danxu@cse.ust.hk, \{ga

Abstract

Multi-task learning for dense prediction has emerged as a pivotal area in computer vision, enabling simultaneous processing of diverse yet interrelated pixel-wise prediction tasks. However, the substantial computational demands of state-of-the-art (SoTA) models often limit their widespread deployment. This paper addresses this challenge by introducing network binarization to compress resource-intensive multi-task dense predictors. Specifically, our goal is to significantly accelerate multi-task dense prediction models via Binary Neural Networks (BNNs) while maintaining and improving model performance at the same time. To reach this goal, we propose a Binary Multi-task Dense Predictor, Bi-MTPD, and several variants of Bi-MTPD, in which a multi-task dense predictor is constructed via specified binarized modules. Our systematical analysis of this predictor reveals that performance drop from binarization is primarily caused by severe information degradation. To address this issue, we introduce a deep information bottleneck layer that enforces representations for downstream tasks satisfying Gaussian distribution in forward propagation. Moreover, we introduce a knowledge distillation mechanism to correct the direction of information flow in backward propagation. Intriguingly, one variant of Bi-MTPD outperforms full-precision (FP) multi-task dense prediction SoTAs, ARTC [2] (CNN-based) and InvPT [50] (ViT-Based). This result indicates that Bi-MTPD is not merely a naive trade-off between performance and efficiency, but is rather a benefit of the redundant information flow thanks to the multi-task architecture. Code is available at BiMTPD.

1. Introduction

There is a growing trend in the computer vision community where dense prediction tasks are processed in a multi-task learning manner, such as semantic segmentation, monocular depth estimation, and human parsing [28, 46, 47, 49].

\textsuperscript{†}Corresponding author
While BNNs have shown impressive results in image classification, achieving nearly full-precision ResNet-level accuracy [25], their application has largely been limited to small-scale models, overlooking other computationally intensive computer vision tasks [15, 23, 25, 32, 39]. Extending BNNs to larger models should be the next step. However, this expansion has been hindered by issues such as overfitting [19, 38, 39] and information degradation [30]. Techniques effective in full-precision models, like label smoothing [34], dropout [42], and mixup [52] have less effect on BNNs [18, 38, 39]. Furthermore, SoTA multitask dense prediction tasks often require deep and complex models, equipped with multi-modality fusion structures [2, 47, 50], exacerbating the challenges in implementing binarization effectively.

The primary barrier to applying binarization in multitask dense prediction tasks is the significant degradation of information flow in deep models [15, 30, 32], leading to reduced performance. To address this issue, we first propose a Binary Multitask Dense Predictor (Bi-MTPD) baseline, where a multi-task dense predictor is formulated via binarized modules. Based on a thorough review of this baseline, we conclude that the binarization operation destroys the information flows in multi-task models, and thus representations for downstream tasks are not informative compared with their full-precision counterparts. To tackle this problem, we update Bi-MTPD with additional information flow calibration mechanisms in two directions. First, we implement variational information bottleneck enforcing the embeddings to follow Gaussian distribution with sparsity in forward propagation, in order to filter out the task-irrelevant factors. Second, we leverage the existing FP models via feature-based knowledge distillation to calibrate the gradient of the binary network in backward pass.

The benefits of Bi-MTPD can be analyzed from two orthogonal perspectives. On one hand, from the perspective of network binarization, the accomplishment of bridging binarization with the multitask dense prediction framework testifies that Bi-MTPD can effectively supplement information, and consequently improve the performance of the individual binary models. On the other hand, from the perspective of multitask dense prediction task, accelerating those cumbersome models is profitable to design more effective and efficient cross-talk modules in it, as shown in Fig. 1. Since existing dense prediction models have severe limitation in modelling the cross-talk modules due to their heavy utilization of convolution operation, it is critical for the multitask dense predictions to learn interactions and inference covering various scopes of the multitask context via the cross-talk mechanism [2, 28, 47, 49, 50, 53]. Intriguingly, a variant of Bi-MTPD outperforms SoTA approach ATRC [2] by 4% over the segmentation task while remaining 43% faster in speed, implying that our proposed method is not a naive trade-off between performance and efficiency. By empirically investigating this “free lunch” achievement, we conclude that the win-win outcome is benefited from our designed information supplementation mechanism which strengthens the representation ability of the binary model.

2. Related Work

Multitask Dense Prediction. Multi-Task Learning (MTL) methods can be generally categorized into two main paradigms in terms of the way where model learns shared representations: hard and soft parameter sharing. Hard parameter sharing characterizes architectures which typically share the first hidden representations among the tasks while branching to independent task-specific representations at a later stage. Most approaches split to task-specific heads at a single branch point [5, 17, 21, 36]. However, such naive branching can be sub-optimal, raising interest in mechanisms that allow for finely branched architectures [26, 45]. As a result, in soft parameter sharing, each task is assigned its own set of parameters and a feature sharing mechanism realize the cross-talk as demonstrated in Fig 1. The following works devise the cross-talk mechanisms focusing on the locations in the network where information or features are exchanged or shared between tasks. Apart from the locations, the feature sharing modules are also widely studied. For example, feature fusing can be introduced along the entire network depth [9, 28]; PAD-Net [49] uses multi-modal distillation to enhance task-specific predictions, in which information flow from each source to target task is regulated with a sigmoid-activated gate function; and MTI-Net [47] combines the multi-modal distillation module of PAD-Net with a multi-scale refinement scheme to facilitate cross-task talk at multiple scales.

Although increasing the number of cross-talk modules intuitively benefits the overall performance of the models, computational cost is often an obstacle. To handle this issue, ATRC [2] introduces NAS [55] to automatically design an efficient information fusing modules. From the perspective of the efficient representation cross-talk, our proposed models with the binarization module can be interpreted as a new pathway to feature fusing within a notably lower inference speed level.

Neural Network Binarization. As pioneers, [15] use the sign function to quantize weights and activations to ±1, initiating the trends of studies of network binarization. To tackle the vanishing gradient issue induced by the binarization operations, the straight-through estimator (STE) [1] is introduced for the gradient approximation. Rooted in this archetype, considerable studies contribute to improving the performance of BNNs, particularly on ImageNet. For example, [23] propose Bi-Real introducing double residual connections with FP downsampling layers to mitigate the excessive gradient vanishing issue caused by binariza-
tion, and consequently demonstrate that delicately placing additional connections within BNNs benefits different propagation. [13] design a proxy matrix as of the latent parameter space to guide the alignment with different bits by recovering the smooth BNNs. In summary, a large number of methods have tended the boundary of the network binarization towards accuracy over classification (e.g., ReActNet [25] in this comparable FLOPs of binary ResNet-18 achieves 65.9% Top-1 accuracy on ImageNet, while full-precision version is only 70.5%).

However, most of those works validate their effectiveness over classification with relatively small architectures (mostly ResNet18 and ResNet34). Meanwhile, the network-based models for dense prediction tasks are bigger and deeper than those toy models, as the information flow in networks is severely degraded. Consequently, directly implementing existing binarization methods can not achieve supposed success. To mitigate the degradation, we propose to binarize those dense prediction models in a multitask way.

3. Multitask Network Binarization

3.1. Binary Neural Network

To begin with, we briefly review the general idea of binary neural networks (BNNs) in [7, 15]. Here, we only elaborate the speedup mechanism and the degradation of information flow of the binarization. We define a full-precision (FP) neural network with $K$ layers, $f(x) = (W^K \times \sigma \times W^{K-1} \cdots \sigma \times W^1)(x)$, where $x$ is the input sample and $W^k : \mathbb{R}^{d_{k-1}} \rightarrow \mathbb{R}^{d_k}$ ($k = 1, ..., K$) stands for the weight matrix connecting the $(k-1)$-th and the $k$-th layer, with $d_{k-1}$ and $d_k$ representing the sizes of the input and output of the $k$-th network layer, respectively. The $\sigma(\cdot)$ function performs element-wise activation for the feature maps.

BNNs vary from FP neural networks in terms of the forward operation and the backward gradient approximation. Specifically, in the forward propagation, the BNN maintains FP latent weights $W_F$ for gradient updates, and the $k$-th weight matrix $W^k_F$ is binarized into $\pm 1$, obtaining the binary weight matrix $W^k_B$ via the binarize function $\text{sign}(\cdot)$, i.e. $W^k_B = \text{sign}(W^k_F)$. Then the intermediate activation map (full-precision) of the $k$-th layer is produced by $A^k_B = W^k_B A^k_B$. The same quantization method is used to binarize the full-precision activation map as $A^k_B = \text{sign}(A^k_F)$, and the whole forward pass of binarization is performed by iterating this process for $L$ times, as shown in Fig. 2. For BNNs, the weights and activations are 1-bit, by which the network is accelerated 32 times in terms of memory cost. Importantly, inference of BNN is accelerated 64 times, as the FP multiplication in FP networks is replaced by Xnor-Bitcount in BNNs.

![Figure 2. A general illustration of the forward propagation of the $k$-th layer in the BNN.](image)

In the backward propagation, the main challenge is that the pervasive sign functions are theoretically non-differentiable, and thus extremely destroy the information flow via gradient the propagation. To address this problem, a large number of researchers [31] widely exploit the straight-through estimator (STE) [1] to numerically approximate the derivative of the whole BNN, i.e.

$$\frac{\partial L}{\partial x} = \begin{cases} \frac{\partial L}{\partial \text{sign}(x)} & |x| \leq 1 \\ 0 & |x| > 1 \end{cases}$$

It is worth noting that we do not implement the aforementioned vanilla approximation method in practice, while we utilize the prevalent Bi-Real [23] and IR-Net [32] to gradually approximate the $\text{sign}$ function, which have been proven to be better estimation approaches [29, 30].

Even though numerous methods have been proposed to eliminate the deterioration of the information flow induced by the binarization, the deterioration is still inevitable due to the severe accuracy loss of weights, activations and gradients [29, 30]. Consequently, binarization destroys the performance of the complicated computer vision models.

3.2. Multitask Dense Predictor

After deploying binarization techniques in the models for dense prediction tasks, the performance of the binarized models is unacceptable, as shown in Fig. 1 and the binary single result in Table. 3. Since the architectures of those SoTA dense prediction models are relatively heavier and deeper (e.g., HRNet-48 or ResNet-101 with a task-specific head [27]) than the ones for classification (e.g., ResNet-18 with a fully-connected layer as the classification head). Moreover, the information passing in the binary models via back-propagation, especially in deep models, is notoriously inefficient [23].

Dense prediction tasks can mutually supplement information, e.g., surface normals and depth can directly be derived from each other, which can be modeled as the regularization of each other [47]. The relevancy among dense prediction tasks is worth being utilized to improve the overall performance of models. For example, before the deep learning era, pioneering work [11] utilizes RGB-D images with depth information to predict scene semantics to improve the quality of the prediction. In the deep learning era, recent attention-based multitask learning methods [2, 47, 49] explicitly and implicitly distill information from other tasks as a complementary signal to improve the targeted task performance. Briefly, the above-mentioned methods are achieved
Multi-Modal Distillation (MMD) via Binarized Attention Mechanism. The multi-modal distillation module is the key in the multi-task dense prediction model. Specifically, we utilize the attention mechanism for guiding the information passing between the binary feature maps generated from different modalities for different tasks. Since the passed information flow is not always helpful, the attention can act as a gate function to control the flow, in other words, to make the network automatically learn to focus or to ignore information from other binary features [47, 49].

Including the binarization operations, we can formalize the MMD via binarized attention as follows. While passing information to the $k$-th task, we first obtain a binarized attention map $A_{B,i}^{k}$, i.e.,

$$A_{B,i}^{k} \leftarrow \text{bool}(W_{B}^{k} \odot F_{B,i}^{k})$$

(2)

where $W_{B}^{k}$ is the parameters of the binarized convolution layer, $F_{B,i}^{k}$ is the binary feature map of the initial prediction, and $\odot$ denotes convolution operation. Then the information is passed with the attention map controlled as follows:

$$F_{B,i}^{o,k} \leftarrow \text{sign} \left[ F_{B,i}^{k} + \sum_{t=1, t \neq k}^{T} A_{B,t}^{k} \odot (W_{B,t} \odot F_{B,i}^{t}) \right]$$

(3)

where $\odot$ element-wise multiplication. The general demonstration of the distillation process is presented in Fig. 3 Left. The output binary feature map $F_{B,i}^{o,k}$ is then used by the head for the corresponding $t$-th task in Fig. 3 Right. By using the task-specific distillation activations, the network can preserve more information for each task [2, 47, 49], which especially benefits the BNNs where the deteriorated information flow mainly induce the performance drop.

On the other hand, multitask dense prediction models benefit from the network binarization in terms of performance. Although these multitask models have achieved a promising performance, they are still limited by the nature of convolution-based distillation modules that are heavily used in a multi-scale way, which model critical spatial and task-related interactions in relatively local perceptive fields [50]. Theoretically, more distillation modules in different network nodes can contribute to model performance, yet we cannot unrestrictedly add distillation modules to the existing model due to the computational limitation. Fortunately, with the saved computational cost of the binary networks, we can implement additional distillation modules in our model.

Binary Baseline for Multitask Dense Prediction, Bi-MTPD. To obtain dense predictions with BNNs under a multitask learning framework, we practically binarize...
the MTI-Net [47] as the binary baseline. Specifically, the main modules in the full-precision MTI-Net, including backbone, heads, and multi-modal distillation module, are replaced with binary modules (both weights and activations are 1-bit). We call this baseline as Bi-MTPD.

3.3. Information Flow Supplementation

Although we build a fully binarized baseline Bi-MTPD for multitask dense predictions and train the pipeline with common techniques, the performance is still of major concern. The baseline suffers an immense information degradation as the nuisance factors are over-fitted in the forward propagation and optimization directions severely mismatch in the backward propagation. To solve these problems, in this section, we further propose the variant of Bi-MTPD, Bi-MTPD-F. Specifically, we introduce a variational information bottleneck (VIB) layer after the output of the shared binary backbone to precisely enforces the feature extractor to preserve the minimal sufficient information of the input data. As well, we deploy the feature-based knowledge distillation mechanism to guide the optimization direction. We present more details in the following section.

Variational Information Bottleneck for Filter-Out Nuisance Factors. Obtaining the initial binary representations of input images by the shared backbone, we need to train a series of targeted heads to split them. A straightforward strategy is to feed these representations into the following MMD module. However, the binarized representations lack homogenization leading to model overfitting issue [48]. Therefore, the need to regularize the binarized representations, while the regularization would not to contaminate the information flow in the representations. Fortunately, the information bottleneck (IB) principle directly relates to compression with the best hypothesis that the data misfit and the model complexity should simultaneously be minimized [43, 48].

As the VIB could effectively capture the relevant parts and filter out the irrelevant ones from inputs, we design a novel VIB-based layer after the backbone. In particular, it explicitly enforces the feature extractor to preserve the minimal sufficient information of the input data. In other words, it can help ensure the information flow flexibly to learn clean representation for the targeted tasks. The objective function of our VIB-based classification can be formulated as a term of information loss, written as follows:

$$\mathcal{L}_{vib} = KL [p(Z \mid A_B), r(Z)],$$  \hspace{1cm} (4)

where $A_B$ is the input binary backbone representation, $B$ is the latent representation variable, $p(Z \mid A_B)$ is a multivariate Gaussian distribution, and $r(Z)$ is a standard normal distribution. Generally, the latter is a regularization term controlling how much information of the input is filtered out. A more detailed discussion about the VIB for binarized models filtering out irrelevant information is in Supplementary Materials.

Feature-based Knowledge Distillation for Guiding the Direction of Information Flow. Distillation is a common and essential optimization approach to alleviate the performance drop of quantized models on ultra-low bit-width settings, which can be flexibly deployed for any architectures to utilize the knowledge of a full-precision teacher model [3, 16, 30, 51]. The usual practice is to distill the activations in a layerwise manner from the full-precision teacher to the quantized counterparts, i.e., $F_{B,l}$ and $F_{FP,l}$ ($l = 1, \cdots, L$, where $L$ represents the number of network layers), respectively. We use the mean squared errors (MSE) as the distance function to measure the difference between corresponding from features student and teacher. The knowledge distillation loss can be written as follows:

$$\mathcal{L}_{kd} = \sum_{l=1}^{L} MSE(F_{B,l}, F_{FP,l}).$$  \hspace{1cm} (5)
3.4. Counter-Intuitive Results of Bi-MTDP-A

Intuitively, implementing binarization on FP network inevitably induces representations degradation, as the gradient of the $\text{sign}$ function cannot be perfectly estimated [1]. Thus, binarized models are impossible to outperform their full-precision counterpart models. However, Bi-MTDP-C, a variant of Bi-MTDP (i.e., full-precision backbone with only binarized multi-modal distillation) outperforms its fully FP version. Specifically, just binarizing the multi-modal distillation can simultaneously accelerate the model by $\sim 39\%$ and improve the mIoU for segmentation by $\sim 4\%$, as shown in Table 1. This result demonstrates that our method is not a naive trade-off between model performance and efficiency but a powerful tool for boosting multitask dense predictors. This exciting ‘free-lunch’ achievement is even a bit of counter-intuitive. We speculate the reasons are that i) binarization on MMD can filter out task-irrelevant information; ii) and thus the information flow in the network is more effective. To testify this speculation, we conduct a series of experiments in two aspects, the representation ability of Bi-MTDP-C and information flow supplementation within the network.

Qualitative Study of Learned Features with Bi-MTDP
To investigate the representation ability of Bi-MTDP-C and its FP counterparts, we visualize i) the feature maps behind the Binarized Multi-Modal Distillation (MMD) module in 2-D space via the t-SNE [44] algorithm, and ii) the regions where the network considers important via the Grad-Cam algorithm [35]. The results are shown in Fig. 5. It is clear that binarized model, Bi-MTDP-C is able to filter the irrelevant information out via the binarized attention module (see Fig. 5 (a)), and thus helps learn more discriminative features (see Fig. 5 (b)) resulting in higher quantitative results. Overall, the generated spatial feature maps for segmentation are better. The enhanced representative ability can contribute to higher quantitative results.

Analysis of Information Flow Supplementation within Network via Centered Kernel Alignment.
Analyzing distributional information flow within layers of neural networks is challenging because outputs of layers are distributed across a large number of neurons. Centered kernel alignment (CKA) [6, 22, 33] can address these challenges, by quantitatively comparing activations within or across networks. Specifically, for a network feed by $m$ samples, CKA algorithm takes $X \in \mathbb{R}^{m \times p_1}$ and $Y \in \mathbb{R}^{m \times p_2}$ as inputs which are output activations of two layers (with $p_1$ and $p_2$ neurons respectively). Letting $K \equiv XX^\top$ and $L \equiv YY^\top$ denote the Gram matrices for the two layers CKA computes:

$$
\text{CKA}(K, L) = \frac{\text{HSIC}(K, L)}{\sqrt{\text{HSIC}(K, K)\text{HSIC}(L, L)}},
$$

where HSIC is the Hilbert-Schmidt independence criterion [10]. Given the centering matrix $H = I_n - \frac{1}{n}11^\top$ and the centered Gram matrices $K' = HKH$ and $L' = HLH$, HSIC $= \frac{\text{vec}(K')\text{vec}(L')}{(m-1)^2}$, the similarity between these centered Gram matrices. Importantly, CKA is invariant to orthogonal transformation of representations (including permutation of neurons), and the normalization term ensures invariance to isotropic scaling. These properties enable meaningful comparison and analysis of neural network hidden representations.

Therefore, we introduce CKA to study the information flow in the multitask dense prediction models. In the heatmap, the lighter the dot, the more similar the two corresponding layers. Higher similar score between two layers’ output representations means those two layers share more information. The results are presented in Fig. 5 (c), we can see that the similar scores among front layers and back lay-
### 4. Experiments

In this section, we conducted comprehensive experiments to evaluate our proposed method on two datasets for dense prediction tasks: PASCAL Context [8] and NYUD-v2 [41]. We first describe the implementation details of Bi-MTPD, and then compare our method with SoTA binary neural networks in the task of object detection to demonstrate superiority of the proposed method. Finally, we validate the effectiveness of information bottleneck and feature-based knowledge distillation by a series of ablative studies.

#### 4.1. Datasets, Evaluation, Implementation Details

**Datasets.** PASCAL-Context is a popular dataset for dense prediction tasks. We use the split from PASCAL-Context which has annotations for semantic segmentation, human part segmentation, semantic edge detection, surface normals prediction and saliency detection. Note that some annotations were distilled by [27] using pre-trained SoTA models [4]. NYUD-v2 contains various indoor scenes such as offices and living rooms with 795 training and 654 testing images. It provides different dense labels, including semantic segmentation, monocular depth estimation, surface normal estimation and object boundary detection.

**Evaluation.** Semantic segmentation (Semseg) and human parsing (Parsing) are evaluated with mean Intersection over Union (mIoU); monocular depth estimation (Depth) is evaluated with Root Mean Square Error (RMSE); surface normal estimation (Normal) is evaluated by the mean error (mErr) of predicted angles; saliency detection (Saliency) is evaluated with maximal F-measure (maxF); object boundary detection (Boundary) is evaluated with the optimal-data-set-scale F-measure (odsF). To evaluate the model efficiency w.r.t. memory cost and inference speed, we adopt the number of parameters and FLOPs for a single round of the model inferring an input image.

**Implementation details.** We build our approach on the most prevalent backbone architecture, i.e., HRNet as previous SoTA methods [2, 47]. The task-specific heads are also implemented as two basic residual blocks, i.e., binarized BasicBlock and binarized Bottleneck with additional binary shortcuts as BiReal-Net [23]. We use $\ell_{1}$ loss for depth estimation and cross-entropy loss for semantic segmentation on NYUD-v2. As in the prior work, the edge detection task is trained with a positive weighted wpos = 0:95 binary cross-entropy loss. We do not adopt a particular loss weighing strategy on NYUD-v2, but simply sum the losses together. On PASCAL, we reuse the training setup from [47] to facilitate a fair comparison. We reuse the loss weights from there. The initial task predictions in the front-end of the network use the same loss weighing as the final task predictions. We refer to the supplementary material for further implementation details. Importantly, we use Adam optimizer [20] in training, but with different learning rates for binary parameters (1e-5) and FP parameters (1e-4), as Adam with a larger learning rate for binary parameters can lead to better training results [24]. Note that our project is based on the codebases for MTI-Net [47] and more details can be found in the codes in the Supplemental Materials.

#### 4.2. Comparison with State-of-the-Art

Tabs. 1 and 2 present a comparative analysis of the proposed Binary Multitask Dense Predictor (Bi-MTPD) with current state-of-the-art models on the NYUD-v2 and PASCAL-Context datasets. This comparison includes notable CNN-based methods such as PAD-Net [49], ASTMT [27], MTI-Net [47], and ATRC [2], among others. Bi-MTPD demonstrates exceptional performance, outperforming other models in 6 of the 9 evaluated metrics, particularly in complex scene understanding tasks like Semantic Segmentation and Parsing. Remarkably, on the NYUD-v2 benchmark, Bi-MTPD surpasses the previously best-performing CNN-based method (ATRC) by a margin of +1.8 (mIoU) in Semantic Segmentation, while requiring...
only 62% of the storage space for weights and 56% of the computational FLOPs.

Furthermore, the application of Bi-MTPD to ViT-based state-of-the-art method, InvPT [50], show cases Bi-MTPD’s ability to enhance model performance while also improving efficiency. This demonstrates the broad applicability and generalization capability of Bi-MTPD across different architectural frameworks.

For a qualitative assessment, Figs. 7 and 6 displays prediction examples from various models. These examples illustrate that Bi-MTPD-C not only competes with but occasionally surpasses the state-of-the-art ATRC in qualitative performance.

4.3. Ablative Studies

Tab. 3 shows a series of ablative studies of Bi-MTPD-C and Bi-MTPD-F with an HRNet-48 backbone on NYUD-v2. We verify how different components of our model contribute to the multi-task improvements. In summary, every designed module positively impacts the overall model. We would like to highlight the main intuition of our method that binarized dense prediction models in a multitask manner largely outperform the binarized single models through Bi-MTPD-F w.o. VIB & KD vs. Bi-Single.

Table 3. Ablative studies on NYUD-v2.

<table>
<thead>
<tr>
<th>Model Method</th>
<th>Backbone</th>
<th>SemSeg ↑</th>
<th>Parsing ↑</th>
<th>Saliency ↑</th>
<th>Normal ↓</th>
<th>Bound ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bi-MTPD-C</td>
<td></td>
<td>47.71</td>
<td>0.530</td>
<td>20.06</td>
<td>77.36</td>
<td></td>
</tr>
<tr>
<td>Bi-MTPD-C w.o. binary MMD</td>
<td></td>
<td>45.70</td>
<td>0.537</td>
<td>20.27</td>
<td>77.86</td>
<td></td>
</tr>
<tr>
<td>Bi-MTPD-F</td>
<td></td>
<td>39.20</td>
<td>0.612</td>
<td>21.04</td>
<td>76.86</td>
<td></td>
</tr>
<tr>
<td>Bi-MTPD-F w.o. VIB</td>
<td></td>
<td>35.04</td>
<td>0.652</td>
<td>22.31</td>
<td>73.88</td>
<td></td>
</tr>
<tr>
<td>Bi-MTPD-F w.o. KD</td>
<td></td>
<td>36.20</td>
<td>0.640</td>
<td>21.99</td>
<td>74.86</td>
<td></td>
</tr>
<tr>
<td>Bi-MTPD-F w.o. VIB &amp; KD</td>
<td></td>
<td>33.99</td>
<td>0.667</td>
<td>23.24</td>
<td>71.30</td>
<td></td>
</tr>
<tr>
<td>Bi-Single</td>
<td></td>
<td>16.20</td>
<td>0.872</td>
<td>29.36</td>
<td>69.24</td>
<td></td>
</tr>
</tbody>
</table>

5. Conclusion

In this paper, we significantly accelerate cumbersome dense prediction models, in which BNNs for relevant tasks are modeled and optimized under a multitask framework to supplement degraded information caused by binarization op-erations. Based on this binary baseline, we further introduce variational information bottleneck and feature-based knowledge distillation to supplement information flow. Experiment results show that our method significantly accelerates existing SoTA methods with comparably small performance drop over the mainstream dense prediction tasks on PASCAL VOC and NYUD-v2. Intriguingly, Bi-MTPD not only reaches SoTA w.r.t. performance but also saves computational costs, compared with SoTA method ARTC [2].

Acknowledgments: This research is partially supported by NSF IIS-2309073, ECCS-212352101 and Cisco unrestricted gift. This article solely reflects the opinions and conclusions of its authors and not the funding agencies.
References


[35] Ozan Sener and Vladlen Koltun. Multi-task learning as


[38] Yuzhang Shang, Dan Xu, Bin Duan, Ziliang Zong, Liqiang Nie, and Yan Yan. Lipschitz continuity retained binary neural network. In ECCV, 2022.


