Vision Perception Unit: Next-Generation Smart CMOS Image Sensor

Wenqi Ji, Yuxing Han, Jiangtao Wen, Yubin Hu, Futang Wang, Yuze He, Xi Li and Jun Zhang

Department of Computer Science and Technology, Tsinghua University
Abstract

As we reach the end of Moore’s Law and Dennard Scaling, it has become highly desirable to design a highly integrated and optimized pipeline specifically for computer vision. A new generation of integrated “smart” visual processors that streamline an end-to-end optimized visual information acquisition and processing pipeline (VIAPP) becomes necessary to lower the cost, power consumption, and latency.

We describe a new paradigm for VIAPP as Vision Perception Unit (VPU), wherein electric signals generated by photons are amplified before converting to the digital signals to emulate an initial layer of a convolutional neural network (CNN). The outputs from these layers are then converted to digital signals and processed by following layers of a deep CNN.
Abstract
Conventional CMOS Image Sensor

- An image signal processor (ISP) for color processing, denoising, correction, etc.
- Further pre-processing in digital signal processor (DSP)
  - e.g., image enhancement and compression
- All capture image frames are processed with original high resolution
• Sensing and processing are integrated
• DSP takes raw images as input
• No delay and power consumption from ISP
• In-pixel filters driven by DSP
VPU: Process and Domain Specific Architecture

Pixels for Dynamic Architecture

- a. Dynamic Resolution
- b. ROI
- c. Dynamic read-out precision

Partially read-out from CMOS Image Sensor
Dynamically lower frequency of ADCs & Clocks,
reduce the bandwidth of data transmission

Vision Tasks

Feedback
Video Segmentation
Edge Extraction
VPU: Sensing-Processing-Integrated Hardware

Stacked Die Package

- Dynamic Pixels
- Backside Illuminated
- VPU Image Processing Algorithms
- DSP Accelerators

Lower power consumption

Increase fill factor

Improve performance of CV applications
Edge Extraction and Video Segmentation on DSP

• The DSP in VPU is optimized for edge extraction and video segmentation in low light for various applications.

• Low light
  • Applied on 24/7 self-driving, AIoT, CCTV, robot, etc.
  • Suitable for high frame rate imaging (~1000fps)
  • Low cost on lens

• Edge extraction and video segmentation
  • Basic feature and semantic label used for other CV tasks
Our Unet-based edge extraction model work on raw images directly from image sensor readout.

Output the contour information for gesture recognition and abnormal behavior detection.

Suitable for extremely low light (20 photons per pixel)
Dynamic Resolution for Video Segmentation

• Processing video frames with dynamic resolution reduces both **read-out cost** and **computation cost**.

**Read-out with Dynamic Resolution utilizing Random Access Ability of CMOS**

**Traditional: Computation with Constant Resolution**

**VPU@DR-Seg: Computation with Dynamic Resolution**
The Cross Resolution Feature Fusion module (CReFF) aggregates HR features into LR features with local attention mechanism.

- A feature similarity loss is designed to aid the training process.
Performance of DR-Seg

- DR-Seg **outperforms** the **state-of-the-art** constant-resolution algorithm by **1.0% mIoU** with only **32.97% FLOPs**.

### Results on CamVid dataset

<table>
<thead>
<tr>
<th>Methods</th>
<th>Resolution</th>
<th>mIoU (%)↑</th>
<th>GFLOPs↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSPNet18 *</td>
<td>1.0x</td>
<td>69.43</td>
<td>309.28</td>
</tr>
<tr>
<td>PSPNet18 *</td>
<td>0.5x</td>
<td>66.87</td>
<td>77.27</td>
</tr>
<tr>
<td>DR\textsuperscript{0.5}-PSP18 (Ours)</td>
<td>1.0x, 0.5x</td>
<td>70.48</td>
<td>101.98</td>
</tr>
<tr>
<td>DR\textsuperscript{0.5}-PSP18 (Ours)</td>
<td>1.0x, 0.3x</td>
<td>69.00</td>
<td>56.33</td>
</tr>
</tbody>
</table>

### Results on Cityscapes dataset

<table>
<thead>
<tr>
<th>Methods</th>
<th>Resolution</th>
<th>mIoU (%)↑</th>
<th>GFLOPs↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSPNet18 *</td>
<td>1.0x</td>
<td>69.00</td>
<td>938.52</td>
</tr>
<tr>
<td>PSPNet18 *</td>
<td>0.5x</td>
<td>63.95</td>
<td>234.63</td>
</tr>
<tr>
<td>DR\textsuperscript{0.5}-PSP18 (Ours)</td>
<td>1.0x, 0.5x</td>
<td>69.03</td>
<td>309.69</td>
</tr>
</tbody>
</table>

We proposed VPU, the next-generation smart CMOS image sensor. VPU pioneeringly integrates image sensing and processing into one chip.

Our results illustrate that the efficiency of video segmentation in VPU is improved by the dynamic-resolution architecture while the accuracy is maintained. The performance of edge detection by VPU outperforms SOTA methods using traditional CIS.

VPU could saves power consumption by end-to-end architecture, which reduces the cost of intermediate processing, and dynamic-resolution algorithms with optimized dynamically controlled pixels, which reduces the cost of computation and read-out.

- Suitable for self-driving and AIoT
- Tape-out in 2023