|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ICCV |  |  |  | ICCV |  |
| #3769 | **ICCV 2021 Submission #3769. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.** | |  | #3769 |  |
|  |  |  |  |
| **000** |  |  |  | **054** |  |
| **001** |  |  |  | **055** |  |
| **002** | Accelerating Vision Task Processing with Sensor RAW Images | |  | **056** |  |
| **003** |  | **057** |  |
|  |  |  |  |
| **004** |  |  |  | **058** |  |
| **005** |  |  |  | **059** |  |
| **006** | Anonymous ICCV submission | |  | **060** |  |
| **007** |  | **061** |  |
|  |  |  |  |
| **008** | Paper ID 3769 | |  | **062** |  |
| **009** |  | **063** |  |
| **010** |  |  |  | **064** |  |
| **011** |  |  |  | **065** |  |
| **012** | Abstract | pression). Subsequently, these RGB images are often pro- | | **066** |  |
| **013** | **067** |  |
|  | cessed by certain CV algorithms for appropriate task exe- | |  |
| **014** | This work explores the feasibility to directly execute | cution (e.g., classification, object detection, segmentation). | | **068** |  |
| **015** | high-level vision tasks using the raw sensor data (i.e., RAW | Such classical RGB-domain CV processing pipeline can be | | **069** |  |
| **016** | image), by which we can completely bypass the Image Sig- | simply represented as |  | **070** |  |
| **017** | nal Processor (ISP) devised in digital cameras for decades. |  |  | **071** |  |
| **018** | y = T (I (xraw)) ; | (1) | **072** |  |
| Due to the lack of large-scale RAW image datasets to train |  |
| **019** | robust RAW-domain vision models, we have developed a |  |  | **073** |  |
| **020** | with I( ) denotes the ISP function, T( ) denotes a specific | | **074** |  |
| generative adversarial network (GAN) to transform existing |  |
| **021** | abundant JPEG compressed RGB images to correspond- | CV task, xraw and y denote the RAW image and task- | | **075** |  |
| **022** | specific outcome, respectively. This conventional paradigm | | **076** |  |
| ing RAW-domain representations. We have also collected |  |
| **023** | has been massively deployed in existing large-scale im- | | **077** |  |
| and labelled a small-scale RAW image dataset dedicated |  |
| **024** | age/video analytic infrastructure, such as event detection | | **078** |  |
| for high-level detection and segmentation tasks. This is the |  |
| **025** | and traffic monitoring [[19](#page9)]. |  | **079** |  |
| first dataset of its kind that can be also used with GAN- |  |  |
| **026** | Recent years have witnessed the explosive growth of CV | | **080** |  |
| based learning to capture the characteristics of real RAW |  |
| **027** | applications. One exciting example is the autonomous cars | | **081** |  |
| images. Experiments demonstrate the competitive efficiency |  |
| **028** | that have multiple video cameras and other sensors (e.g., | | **082** |  |
| of RAW-domain object detection and semantic segmentation |  |
| **029** | ultrasonic, LiDAR, and radar) equipped for event detec- | | **083** |  |
| over the conventional paradigm using RGB images. In ad- |  |
| **030** | tion, recognition, and decision-making. Although conven- | | **084** |  |
| dition, performing vision tasks using RAW images removes |  |
| **031** | tional RGB-domain CV engines still dominate in practice, | | **085** |  |
| the need for ISP, which makes it attractive for devices and |  |
| **032** | they pose critical challenges for ultra low-latency | appli- | **086** |  |
| applications with stringent power and latency constraints. |  |
| **033** | cations since delay is inevitably introduced by the serial | | **087** |  |
| We will make our dataset and relevant material publicly ac- |  |
| **034** | transformations in the ISP module [[13](#page9)]. It is generally re- | | **088** |  |
| cessible for reproducible research. |  |
| **035** | ported that the autonomous cars generally require less than | | **089** |  |
|  |  |
| **036** |  | 0.1s for prompt event response. Taking an On-semi ISP ( | | **090** |  |
| **037** | 1. Introduction | <https://www.onsemi.com/>) chip as an example, it | | **091** |  |
| **038** | requires approximately 0.033s to transform an input RAW | | **092** |  |
|  |  |
| **039** | Computer vision (CV) enables high-level understanding | image to its corresponding RGB representation, which is | | **093** |  |
| **040** | of digital images and videos that are mostly processed in the | nearly 33% of the response time and imposes a critical im- | | **094** |  |
| **041** | RGB format. This is probably because many existing CV | pact in applications. |  | **095** |  |
| **042** | algorithms mimic the cortical process of human visual sys- | Aforementioned issue can be largely resolved if we re- | | **096** |  |
| **043** | tem (HVS) to perceive the colors (e.g., red/green/blue light | move the ISP subsystem by directly processing sensor RAW | | **097** |  |
| **044** | radiance sensation of retinal cells) and understand the scene | images for tasks. This raises a fundamental question: can | | **098** |  |
| **045** | for decision making [[15](#page9)]. Digital cameras have separated | we maintain the same vision task efficiency in RAW do- | | **099** |  |
| **046** | the imaging sensor and the ISP to enable different function- | main as in the conventional RGB domain? To address this | | **100** |  |
| **047** | ality [[3](#page9)], where a standard CMOS or CCD sensor first ag- | question, this paper attempts to accelerate the execution of | | **101** |  |
| **048** | gregates light photons as the RAW image to characterize | vision tasks using sensor RAW images directly. We test our | | **102** |  |
| **049** | the photometric attributes of captured scene, and then the | proposed methods on two prevalent CV tasks, namely ob- | | **103** |  |
| **050** | Image Signal Processor (ISP) converts input RAW image | ject detection and semantic segmentation, to demonstrate | | **104** |  |
| **051** | into corresponding HVS perceivable RGB image through | the efficiency and competitive accuracy compared to con- | | **105** |  |
| **052** | a series of linear and nonlinear steps (e.g., demosaicing, | ventional methods using RGB images converted by ISP. | | **106** |  |
| **053** | white balance, exposure control, gamma correction, com- | Marginal accuracy drop (nearly 1%) may come from the | | **107** |  |

1

ICCV

#3769

**ICCV 2021 Submission #3769. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.**

ICCV

#3769

1. fact that CV algorithms used for demonstration have been
2. extensively optimized for RGB content rather than the RAW
3. data. We expect better performance if these CV algorithms
4. can be refined using a large number of RAW samples. This
5. is deferred as our future study.

**113**

In order to perform RAW-domain tasks, we first col-

**114**

lected 1153 RAW images using iPhone XS Max, which we

**115**

call iPhone RAW Scape 1k (iRAW). The particular focus

**116**

of the collected data was on autonomous car driving appli-

**117**

cations. This dataset is later manually labelled with detec-

**118**

tion bounding boxes and segmentation cues with the help

**119**

of a third-party professional image labelling firm. Nearly

**120**

1000 samples are insufficient for training a robust neural

**121**

model in practice; therefore, the iRAW dataset is primar-

**122**

ily used for testing/inferenc and knowledge transferring in

**123**

RAW-domain fine-tuning .

**124**

Even though we have several image datasets (e.g., Im-

**125**

ageNet) for various vision tasks, they are stored in RGB

**126**

format and cannot be used for RAW processing directly.

**127**

Given the nonlinear transformations involved in camera ISP,

**128**

we propose to approximate the inverse functions in cam-

**129**

era ISP, noted as “invISP”, and another mirroring “neural-

**130**

ISP” as the reverse process of invISP, of which we follow

**131**

the CycleGAN [[23](#page9)] architecture to well capture the cross-

**132**

domain characteristics. Since most RGB images are JPEG

**133**

compressed for storage, we also include the compression

**134**

artifacts removal and quality estimation into the training to

**135**

improve the model robustness. For simplicity, we use the

**136**

term “JPEG” or simply “RGB” to represent decoded RGB

**137**

image with compression noise.

**138**

In the end, the GAN-based invISP works in an unsuper-

HRNetv2 [[17](#page9)] for semantic segmentation. Experimental re-sults show that our method achieves comparable accuracy to the native solutions that use RGB images, while reduc-ing latency and power consumption since our RAW-domain processing completely bypasses the ISP.

2. Related Work

This section briefly reviews the techniques involved in traditional ISP, and RAW image processing.

2.1. Conventional ISP Basics

A serial processing steps have been involved in modern ISPs to convert sensor RAW data to final RGB represen-tation, as shown in Figure [1](#page3)(a). First, linear transforma-tions including the demosaicing, white balance estimation, color contrast adjustment, etc., are applied to convert the native input RAW data to image format conforming to the CIE 1931 XYZ color space [[9](#page9)]. Subsequently, nonlinear transforms are facilitated to translate the image from the XYZ domain to RGB color space, mimicking the nonlinear processing capabilities of the HVS. Well-known nonlinear operations include the denoising, exposure control, gamma correction. Then, JPEG or MPEG based coder is used to reduce the size of RGB images for storage or transmission. This processing pipeline is widely used in modern cameras as the ISP subsystem.A brief white paper on ISP system can also be found here [2](#page2).

2.2. RAW Image Processing

Though most algorithms for image processing have been

**162**

**163**

**164**

**165**

**166**

**167**

**168**

**169**

**170**

**171**

**172**

**173**

**174**

**175**

**176**

**177**

**178**

**179**

**180**

**181**

**182**

**183**

**184**

**185**

**186**

**187**

**188**

**189**

**190**

**191**

**192**

**139**

**140**

**141**

**142**

**143**

**144**

**145**

**146**

**147**

vised manner by mapping an RGB source into its corre-sponding RAW image, while using randomly selected real-life RAW images from in iRAW as a discriminator to effi-ciently captures the cross-domain (e.g., RGB to RAW) char-acteristics for a target camera. Such a GAN-based invISP demonstrates an encouraging efficiency in processing and can potentially help resolve the shortage of RAW images for model training.

Contributions.

applied on RGB images, recent years have shown a number of explorations on RAW image processing [[10](#page9), [2](#page9), [21](#page9), [1](#page9)]. For instance, Zamir et al. [[21](#page9)] proposed a CycleISP to re-map the RGB image to its RAW presentation for denois-ing, by which better image quality was achieved than that in RGB domain. Recently, Afifi et al. [[2](#page9)] demonstrated bet-ter results for various low-level tasks including denoising, color correction, super-resolution, etc., using RAW images that were reversely restored from the input RGB samples.

**193**

**194**

**195**

**196**

**197**

**198**

**199**

**200**

**201**

**148**

1) To the best of our knowledge, this iRAW is the first

**149**

sensor RAW image dataset with high-level detection and

**150**

segmentation labels, which is different from existing sen-

**151**

sor RAW images used for low-level denoising [[21](#page9), [2](#page9)] and

**152** KITTI raw images[1](#page2) (uncompressed RGB images). We will

**153**

continue acquiring RAW images from various scenarios and

**154**

camera models to expand this iRAW dataset.

**155**

These studies have promised encouraging prospect of RAW imaging processing, arising the desire of original RAW im-ages.

Since there is not sufficient real-life RAW images to train CV models, a straightforward way is to employ the inverse ISP functions to reversely transform the exist-ing abundant RGB images into corresponding RAW sam-ples [[10](#page9), [21](#page9), [2](#page9), [4](#page9), [14](#page9)]. But most algorithms like “reverse

**202**

**203**

**204**

**205**

**206**

**207**

**208**

**209**

**156**

2) Our proposed GAN-based invISP simulates the in-

imaging pipeline” (a.k.a., InvGamma) [[10](#page9)] only utilized the

**210**

**157**

**158**

**159**

verse ISP function in an unsupervised fashion, which helps us generate sufficient training samples using existing RGB dataset to retrain the YOLOv5 [[16](#page9)] for object detection and

inverse gamma correction to simulate the nonlinear func-tion and overlooked other important nonlinear processing

2 [https://www.pathpartnertech.com/camera-tunin](https://www.pathpartnertech.com/camera-tuning-understanding-the-image-signal-processor-and-isp-tuning/)

**211**

**212**

**213**

1. 1 [http://www.cvlibs.net/datasets/kitti/raw\_data](http://www.cvlibs.net/datasets/kitti/raw_data.php)
2. [.php](http://www.cvlibs.net/datasets/kitti/raw_data.php)

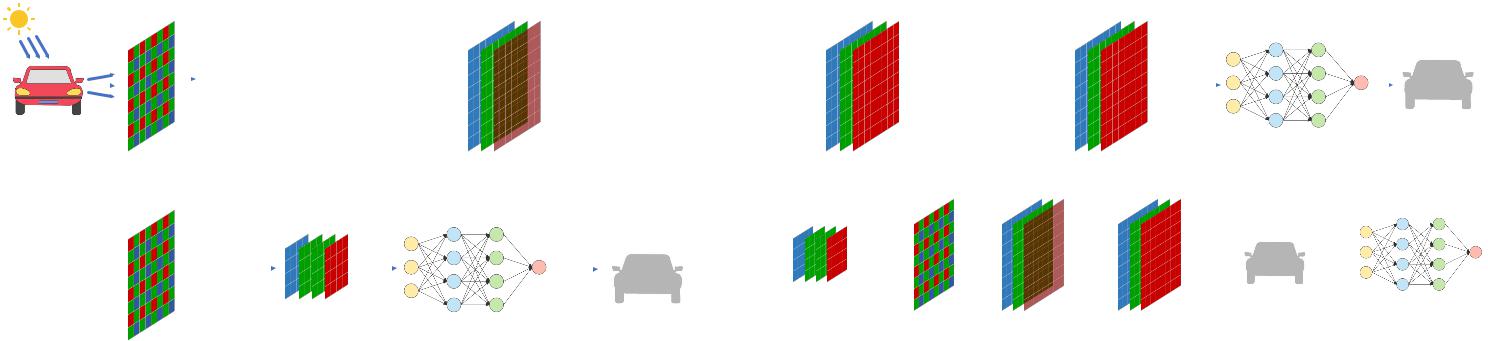
[g-understanding-the-image-signal-processor-and-i](https://www.pathpartnertech.com/camera-tuning-understanding-the-image-signal-processor-and-isp-tuning/) [sp-tuning/](https://www.pathpartnertech.com/camera-tuning-understanding-the-image-signal-processor-and-isp-tuning/)

**214**

**215**

2

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ICCV | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | ICCV |  |
| #3769 |  |  |  |  |  |  |  |  |  | **ICCV 2021 Submission #3769. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  | #3769 |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **216** |  | | | a) |  | | | |  | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  | | | | | | | | | | | | | | | | **270** |  |
|  |  |  |  |  |  |  |  | Image Signal Processing (ISP) | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **217** |  | | | | | | | |  | | | | | |  | | | | |  | |  | |  |  | |  | | | | | |  | | | |  | | | | | | | | |  | | | | | | | | | | | | | | | | **271** |  |
|  |  |  |  |  |  |  |  | Linearization | | | | | | Demosaicing | | | | |  |  |  | Denoising | | Exposure |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | car 0.6 |  | |  |
| **218** | | | |  |  |  | |  |  | | | | | |  | | | | |  | |  | |  |  | |  | | | | | | JPEG | | | |  | | | | | | | | |  | |  | | | | | | | | |  |  |  | | **272** | |  |
|  |  |  |  |  |  |  |  |  |  | White | | |  |  | Color Contrast | | | | |  |  | Sharpening | | | Gamma |  |  |  |  |  |  |  | Compression | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **219** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **273** |  |
|  |  |  |  |  |  |  |  | Balance | | | | |  |  |  | Matrix | | |  |  | Correction |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | | | | | | | |  |  | |  | |  | |  |  |  | | | | | |  | | | |  | | | | | | | | |  | | | | | | | | | | | | | | | |  |
| **220** |  |  |  |  |  |  |  |  |  | Linear Transformations | | | | | | | | | |  |  | Non-linear Transformations | | | |  |  |  |  |  |  |  | Compression | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **274** |  |
| **221** | | | | | | | | |  | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **275** | | | | | | | | | | | | | | | | |  |
| **222** |  | | | b) |  | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **276** |  |
|  | | |  | | | | | | | | | | | | | | | | | | | car 0.6 |  | | | | | | | | | | | | | | | | | | | | | | | | | car 0.6 | |  |  | | | | | | | | |  |
| **223** | | | | | | | rearrange | | | |  | | | | | |  | |  | |  | |  |  |  | | | | | | | | | | | | | | | | | | | | | | | | |  | | | **277** | | | | | | | | | |  |
| **224** |  | | | | | | | | | | | | | | | | | | | | | | |  |  | | | | | | | | | | | | | | | | | | | | | | | | |  | | |  | | | | | | | | | **278** |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | Rearrange | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | CV Tasks | | | |  |  |  |  |  | Neural | | |  |
| **225** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | Sensor RAW | | | CIE-XYZ | | |  |  |  |  | RGB | | |  |  |  |  |  |  |  |  |  |  | **279** |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | RAW | | |  |  |  |  |  |  |  |  |  |  |  | Output | | | |  |  |  |  | Network | | | |  |
| **226** | Figure 1. Camera Pipeline for High-Level Vision Tasks. (a) Traditional RGB-domain CV from senor RAW imaging, ISP processing to | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **280** |  |
| **227** | task execution on RGB samples; (b) Proposed RAW-domain CV by removal of ISP and task execution on RAW images directly. | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  |  |  | **281** |  |
| **228** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **282** |  |
| **229** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **283** |  |
| **230** | in camera ISP. With the help of deep learning techniques, | | | | | | | | | | | | | | | | | | | | | | | | | datasets that have been captured by different cameras do | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **284** |  |
| **231** | **285** |  |
| CycleISP [[21](#page9)] and CIE-XYZ Net [[2](#page9)] suggested to learn in- | | | | | | | | | | | | | | | | | | | | | | | | | not offer explicit labels of their ISPs. Therefore, it is diffi- | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  |
| **232** | **286** |  |
| verse ISP functions through a supervised approach. As a | | | | | | | | | | | | | | | | | | | | | | | | | cult to map these RGB images back to their original RAW | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  |
| **233** | **287** |  |
| result, a large amount of RAW and RGB image pairs are | | | | | | | | | | | | | | | | | | | | | | | | | data. Without the loss of generality, let s and t represent the | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  |
| **234** | **288** |  |
| required to supervise the learning and the learnt model will | | | | | | | | | | | | | | | | | | | | | | | | | source camera for capturing RGB datasets and the target | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  |
| **235** | **289** |  |
| be camera ISP dependent, making it difficult to generalize | | | | | | | | | | | | | | | | | | | | | | | | | camera for inferring the RAW samples, respectively. The | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  |
| **236** | **290** |  |
| itself to various camera ISP models. | | | | | | | | | | | | | | | | | | | | |  |  |  |  | process of reconstructing the RAW image from a JPEG im- | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  |
| **237** |  |  |  |  | **291** |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | age for a target camera can be written as | | | | | | | | | | | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |
| **238** | 3. Methods | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **292** |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **239** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x~rawt = L 1 | | | | |  | N 1 | | | | Cs | | |  | t J 1 (xjpegs ) | | | | | | | | | | | : | | | (3) | | | **293** |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | ! |  |
| **240** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **294** |  |
|  |  | This section describes our exploration in RAW domain | | | | | | | | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  |  | t |  |  | t | |  |  |  |  |  | |  |  |  |  |  |  |  |  |  |  | | |  |  |  |  |
| **241** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **295** |  |
| **242** | to perform the vision tasks, the dataset generation, and the | | | | | | | | | | | | | | | | | | | | | | | | | Cs!t( ) describes cross-domain adaptation from the RGB | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **296** |  |
| inverse ISP development. | | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  | image of source camera to that of the target camera. x~raw | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  |
| **243** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | denotes a simulated version of x | | | | | | | | | | | | | | | | | | | | raw | | | . | | Both | | | | Cs!t | | | | ( ) and | | | **297** |  |
| **244** | 3.1. Problem Formulation | | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  | 1 | ( | ) | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | | **298** |  |
|  |  |  |  |  |  |  |  |  |  | are non-linear operations. For | | | | | | | | | | | | | | | | | simplicity, we merge | | | | | | | | | | | | |  |
| **245** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | Nt | |  |  |  |  |  |  |  | 1 |  |  |  |  |  |  |  |  |  | **299** |  |
|  |  | As illustrated in Figure [1](#page3)(a), the conversion of a RAW | | | | | | | | | | | | | | | | | | | | | | | them using a single nonlinear function N | | | | | | | | | | | | | | | | | | | | | | | | | |  |  | : |  |  |  |  |  |  |  |  |
| **246** | image xraw 2 Rh w to JPEG coded RGB representation[3](#page3) | | | | | | | | | | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  | 1 | |  |  |  | 1 |  |  |  |  | 1 |  |  |  | t!s | | |  |  |  |  |  |  |  |  | **300** |  |
| **247** |  |  |  |  |  |  | x~rawt | | | = Lt | |  | Nt | |  |  | J | | (xjpegs ) | | | | | | |  | : |  |  | (4) | | | **301** |  |
| x | | jpeg 2 | | Rh w 3 can be divided into three steps: linear | | | | | | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  |  | s | | |  |  |  |  |  |
| **248** |  | |  | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  | ! | | | |  |  | | | | | | | | | |  |  | | | | | | **302** |  |
| transformation, nonlinear transformation, and JPEG com- | | | | | | | | | | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **249** | pression. | | | | Here h and w represent the image height and | | | | | | | | | | | | | | | | | | | | |  | We can use the function in ([4](#page3)) to map any RGB image | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **303** |  |
| **250** | to its corresponding RAW representation, and train a task | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **304** |  |
| width, respectively. | | | | | | | | | We can represent mapping between | | | | | | | | | | | | | | | |  |
| **251** | model T optimal | | | | | | | | | | that processes RAW images in infer- | | | | | | | | | | | | | | | | | | | | | | | | | | **305** |  |
| xraw and xjpeg as the following steps: | | | | | | | | | | | | | | | | | | | | |  |  |  |  |  |
| **252** |  |  |  |  | ence, but uses | | | | | | | | RGB images in training, i.e., | | | | | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  | **306** |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | | | | | | | | | | | | | | | | | | | | | | | | |  | | | | | | | | | | | | | | | | | | | | | | | | | | | |  |
| **253** |  |  | xxyz = L (xraw) ; xrgb = N (xxyz) ; xjpeg = J (xrgb) ; | | | | | | | | | | | | | | | | | | | | | | |  |  |  |  |  |  | y~ = T ; L 1 | | | | | | |  | N 1 | | |  |  | J 1 (xjpegs ) | | | | | | | | | | | : | |  | (5) | | | **307** |  |
| **254** |  |  |  |  |  |  |  |  |  | s | |  | **308** |  |
| **255** | where L is a linear transformation from RAW to CIE-XYZ | | | | | | | | | | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | ! | |  |  | |  |  |  |  |  |  |  |  |  |  | | | |  |  |  | **309** |  |
| **256** | image xxyz 2 Rh w 3, N is a nonlinear trasnform that | | | | | | | | | | | | | | | | | | | | | | | | | Ideally we can optimize the network parameters optimal by | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **310** |  |
| solving the following optimization problem: | | | | | | | | | | | | | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |
| **257** | maps CIE-XYZ to RGB image, and J denotes the JPEG | | | | | | | | | | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **311** |  |
| **258** | compression to reduce the size of native RGB images for | | | | | | | | | | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  | optimal = arg min ky~ yk : | | | | | | | | | | | | | | | | | | | |  |  |  |  |  | (6) | | | **312** |  |
| **259** | efficient storage and transmission. Since different cameras | | | | | | | | | | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  | **313** |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | optimal | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **260** | can have different types of sensors and ISPs, the processing | | | | | | | | | | | | | | | | | | | | | | | | | The inference process after the training can be described as | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **314** |  |
| **261** | pipeline for a specific camera c can be represented as | | | | | | | | | | | | | | | | | | | | | | | |  | **315** |  |
| **262** | | | | | | | | | | | | | J |  | N |  | | L |  | | | | | | | | | | | | | | | | | | |  | |  | | | | | | | | | | |  | **316** | | | | | | | | | | |  |
| **264** | | | | | | | | | | | | |  |  | |  | | | | | | | | | | | | | | | | | | |  | | | | | | | | | | | **318** | | | | | | | | | | |  |
| **263** |  |  |  |  | xjpegc | | | | | = | |  |  | ( |  | c ( | |  | c (xrawc ))) : | | |  |  |  | (2) |  |  |  |  |  |  |  |  |  | y~ = T optimal; xrawt | | | | | | | | | | | | | | | | : |  |  |  |  |  |  |  | (7) | | | **317** |  |
| **265** |  |  | In practice, RAW image acquisition is dependent on | | | | | | | | | | | | | | | | | | | | | | | 3.2. Learned invISP | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **319** |  |
| camera/ISP; ideally we would like to assure the same cam- | | | | | | | | | | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **266** |  | We propose a CycleGAN [[23](#page9)]-based approach for self- | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **320** |  |
| era/ISP for the training and testing. However, existing RGB | | | | | | | | | | | | | | | | | | | | | | | | |  |  |
| **267** | supervised domain transform. CycleGAN [[23](#page9)] requires two | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **321** |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **268** |  |  |  | | | | | | | | |  | | | | | | | | | | | | | | generators for source-to-target and target-to-source domain | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **322** |  |
|  |  | 3We exemplify our method using JPEG because most RGB datasets are | | | | | | | | | | | | | | | | | | | | | | |  |
| **269** | coded using popular JPEG standard. | | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  | transformations, and two discriminators (one each in the | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **323** |  |



3

ICCV

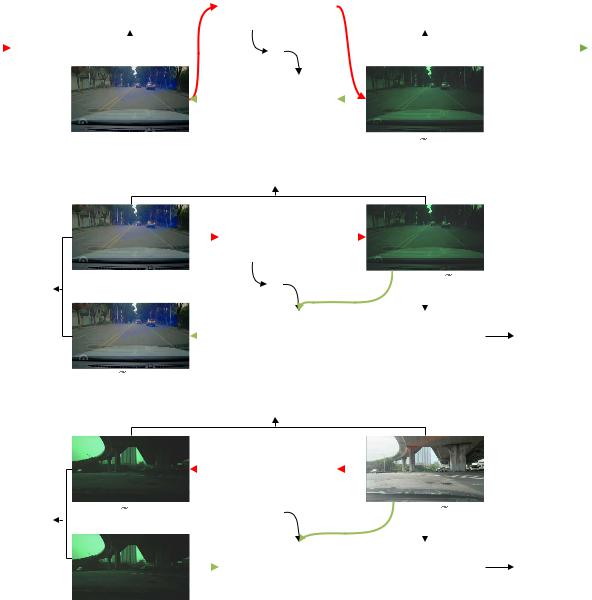
#3769

**ICCV 2021 Submission #3769. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.**

ICCV

#3769

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **324** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  | DJPEG | | |  |  | invISP |  |  | DRAW | |  |  |  |  |  |  |
| **325** |  |  |  |  | |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  | *Q* |  |  |  |  |  |  |  |  |  |  |
| **326** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| JPEG to RAW | | |  |  |  |  |  |  |  |  |  |  |  | RAW to JPEG | | |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| **327** |  |  |  |  |  |  |  |  | neuralISP |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **328** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  | *x jpeg* | |  |  |  | (a) |  |  | *xraw* | |  |  |  |  |  |  |
| **329** |  |  |  | *s* | | |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  | *t* |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  | content loss |  |  |  |  |  |  |  |  |  |  |
| **330** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **331** | cycle- | | |  |  |  |  |  | invISP |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **332** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| consistency | | | *x* | *jpeg* | *s* | | | *Q* |  |  |  | *x* | |  |  |  |  |  |
|  |  |  |  |  |  |  | *t* |  |  |  |  |  |
| **333** |  |  |  |  |  |  |  |  |  |  |  |  | *raw* |  |  |  |  |  |
| loss | | |  |  |  |  |  | neuralISP |  |  | DRAW | |  |  | discriminat | | |  |
| **334** |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  | or loss | | |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **335** |  |  |  |  |  | |  | | (b) |  |  |  |  |  |  |  |  |  |  |
|  |  |  | *x jpegs* | | |  |  |  |  |  |  |  |  |  |  |
| **336** |  |  |  |  |  |  |  |  | content loss |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **337** |  |  |  |  |  |  |  |  | invISP |  |  |  |  |  |  |  |  |  |  |
| **338** | cycle- | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  | *Q* |  |  |  |  |  |  |  |  |  |  |
| **339** | consistency | | | *x* | *t* |  |  |  |  |  |  | *x* | *jpeg* | *s* | | | |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  | *raw* |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **340** | loss | | |  |  |  |  |  | neuralISP |  |  | DJPEG | |  |  | discriminat | | |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  | or loss | | |  |
| **341** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  | *xrawt* | |  |  |  | (c) |  |  |  |  |  |  |  |  |  |  |
| **342** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **343** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |



1. Figure 2. The GAN-based invISP: (a) Our model follows the Cy-

cleGAN [[23](#page9)] architecture, having respective JPEG to RAW map-

**345**

ping and RAW to JPEG mapping, as well as associated adver-

**346**

sarial discriminators Draw and Djpeg. Cycle consistency loss is

1. used to assure the cross-domain consistency, discriminator loss
2. forces generated image to the real image as close as possible,
3. and content loss is to maximize the context similarity between
4. input source and generated output. The quality factor Q esti-
5. mated from JPEG is used to help neuralISP adjust the compres-
6. sion level. (b) Forward pipeline start from JPEG: xjpegs !
7. invISP (~xjpegs ) ! xrawt ! neuralISP (~xrawt ) ! x~jpegs . (c)
8. Backward pipeline start from RAW: xrawt ! invISP (xrawt ) !
9. x~jpegs ! neuralISP (~xjpegs ) ! x~raws . The Q in the backward

process remains the same as the forward process. It is worth men-

**356**

tioning that the xraw mentioned in our paper is bilinearly upsam-

**357**

pled from mosaic xraw mosaic to maintain a consistent resolution

1. with xrgb.

**360**

**361** source and target domain) to distinguish between real and

**362** generated images. In particular, as shown in Figure [2](#page4), we

**363** need an invISP to convert JPEG coded RGB image to its

**364** RAW format, a neuralISP to convert a RAW image to its

**365** JPEG coded representation, one discriminators to distin-

**366** guish between xjpegs and x~jpegs , and the other discrimi-

**367** nator to distinguish between xrawt and x~rawt . To make the

**368** generated JPEG images closer to the original JPEG images,

**369** we use a quality estimation network to estimate the quality

**370** factor Q.

**371** Following the modular decomposition in ([4](#page3)), we have

**372** five sub-networks: linear transformation network, non-

**373** linear transformation network, JPEG artifacts removal net-

**374** work, discriminator, and quality estimation network, as il-

**375** lustrated in Fig. [3](#page5) to fulfill the invISP and neuralISP under

**376** the CycleGAN framework.

**377** Our linear transformation network is inherited from the

CIE-XYZ Net [[2](#page9)], having input at a fixed size of 128 128. It includes five blocks of 3 3 conv - LReLU - 2 2 max pooling layers. Different from CIE-XYZ Net [[2](#page9)], we use reflection padding instead of zero padding to avoid infor-mation loss. The max pooling layers have a stride factor of 2 used for down-sampling. Then FCN with 18 output is used to formulate 3x6 color matrix to perform a linear transformation.

Our nonlinear transformation network and quality es-timation network are modified from CA-Unet [[7](#page9), [12](#page9)], which consists of channel attention-convolutions (CA-Convs) blocks. CA-Convs block first uses global pooling to extract spatial information from convolutional features, and then transforms them via fully connected (FC) layers, ReLU, and sigmoid. Finally, it multiplies the convolutional features with sigmoid’s output, which represent the channel attention’s weights. For nonlinear transformation network, the number of output channels is set to 6 to interface with the linear transform network. While for quality estimation network, we added the average pooling layer after the out-put layer with only one channel to perform the compression quality factor estimation.

The JPEG artifacts removal network is adopted from QGCN [[11](#page9)], which consists of restoration branch and global branch. The restoration branch consists of several resid-ual blocks for extracting the local features and the global branch is used to extract the global features. Both of them are merged together in the middle of the restoration branch. Different with QGCN, we drop out the input qtables be-cause we found that the network can adjust the noise reduc-tion adaptively without applying prior knowledge of qta-bles.

Finally, our discriminator directly adopts the widely used SA-GAN [[22](#page9)], which allows attention-driven, long-range dependency modeling for image generation tasks. We be-lieve that self-attention can be paired with our nonlinear transform network to treat different objects in the scene dif-ferently.

3.3. RAW-domain Vision Tasks

After obtaining the invISP, we can convert xjpegs from any JPEG dataset to simulated RAW image x~rawt to train RAW-domain vision models.

Note that image resolution impacts the the computer vi-sion task. This is because if the resolution is not sufficient large, small objects may be missed by the algorithm. There-fore, for the detection and segmentation tasks, we integrate the upsampling of RAW data re-arrangement for algorithm input, e.g., x~upt = bilinear upsample(mosaic(~xrawt )). In this work, we simply use a bilinear filter to avoid any po-tential performance impacts from the advanced upsampling filters, particularly for those learnt approaches. This also ensures the fair comparison of task-oriented algorithms on

**378**

**379**

**380**

**381**

**382**

**383**

**384**

**385**

**386**

**387**

**388**

**389**

**390**

**391**

**392**

**393**

**394**

**395**

**396**

**397**

**398**

**399**

**400**

**401**

**402**

**403**

**404**

**405**

**406**

**407**

**408**

**409**

**410**

**411**

**412**

**413**

**414**

**415**

**416**

**417**

**418**

**419**

**420**

**421**

**422**

**423**

**424**

**425**

**426**

**427**

**428**

**429**

**430**

**431**

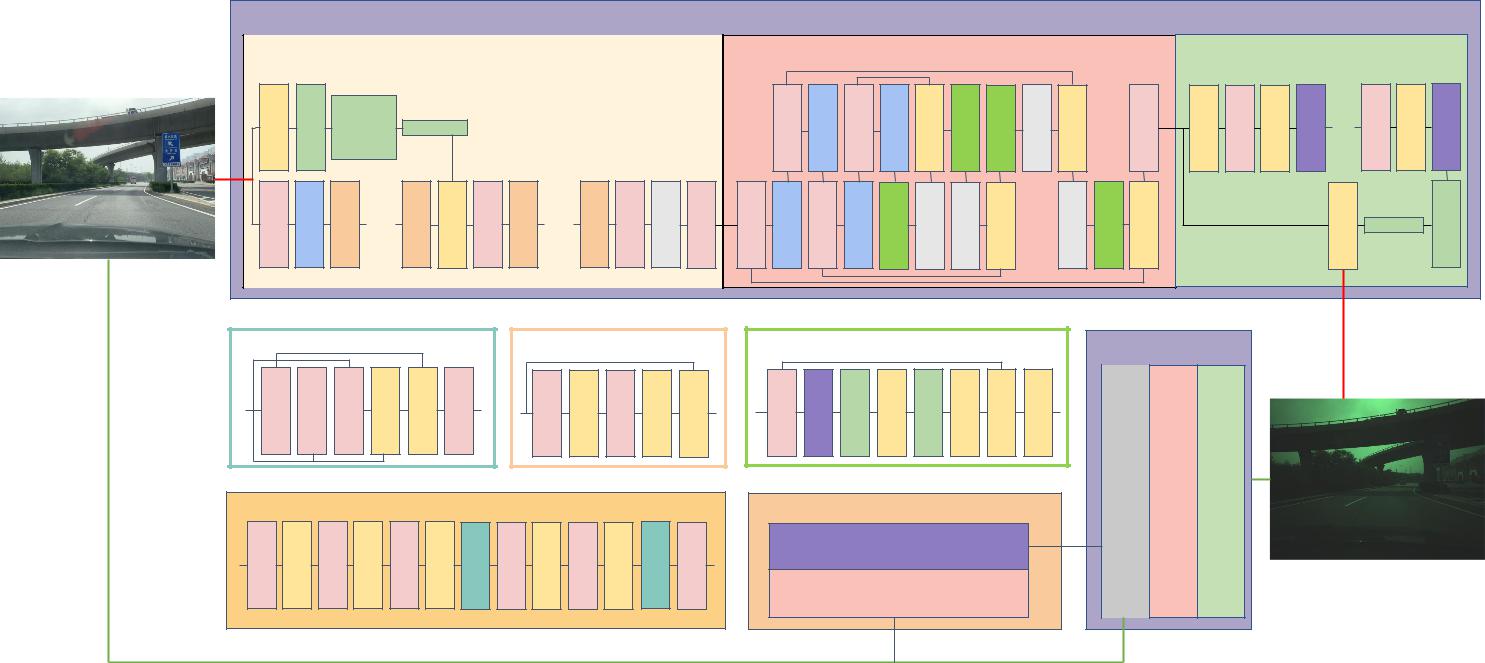
4

ICCV ICCV

#3769 #3769

**ICCV 2021 Submission #3769. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **432** | invISP | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **486** |  |
| **433** | JPEG Artifacts Removal Network | | | | | | |  |  |  |  |  | Non-linear Transformation Network (CA-Unet) | | | | | | | |  |  |  | Linear Transformation Network | | | | | | **487** |  |
| **434** |  | Resize | FC |  | FC |  |  |  |  |  |  |  | Conv | Down | Conv | Down | Concat | Block-CA | Block-CA | UPConcat |  |  | Conv | Resize | ConvLReLU | MaxPool |  | ConvLReLU | MaxPool | **488** |  |
| **436** |  |  | CF |  |  |  |  |  |  |  |  |  | **490** |  |
| **435** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | … |  |  | **489** |  |
| **437** |  | Conv | Down | ResBlock |  | ResBlockConcatInception | ResBlock |  | ResBlock | Conv | UP | Conv | ConvDown | Conv | Down | Block-CA | UP | UP | Concat | UP | Block-CA | | Concat |  |  |  | Mul |  | FC | **491** |  |
| **438** |  | … | … |  |  |  | CF | **492** |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **439** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **493** |  |
| **440** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **494** |  |
| **441** | SA-Block | | |  |  |  | ResBlock | | |  |  |  | CA-Block | |  |  |  |  |  |  | neuralISP | | |  |  |  |  |  |  | **495** |  |
| **442** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **496** |  |
|  | Conv | LReLU | ConvConv | LReLUMul | ConvMulLReLUConvBlock-SA | Conv | LReLUConv | ConvReLU | LReLU | Block-SA | Conv |  | GloPool |  |  |  | Sigmoid | Mul | LReLU | Compressor | | JPEGDifferentiableNetwork | Transformationlinear-Non | NetworkTransformationLinear |  |  |  |  |  |
|  |  | Conv | FC | ReLU | FC |  |  |  |  |  |  |
| **443** |  | Conv | Conv |  |  |  |  |  |  | Conv | Mul | Concat |  |  |  |  |  |  |  |  |  |  |  | **497** |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **444** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **498** |  |
| **445** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **499** |  |
| **446** | Discriminator | | | |  |  |  |  |  |  |  |  | Quality Estimation Network | | | | | |  |  |  |  |  |  |  |  |  |  |  | **500** |  |
| **447** |  |  |  |  |  |  |  |  |  |  |  |  |  |  | Avg-Pool | | |  |  |  |  |  |  |  |  |  |  |  |  | **501** |  |
| **448** |  |  |  |  |  |  |  |  |  |  |  |  |  |  | CA-Unet | | |  |  |  |  |  |  |  |  |  |  |  |  | **502** |  |
| **449** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **503** |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **450** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **504** |  |
| **451** | Figure 3. The architecture of our invISP and the corresponding discriminator, neuralISP. The red line indicates the process from JPEG to | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **505** |  |
| **452** | RAW, the green line indicates the process from RAW to JPEG. | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **506** |  |
| **453** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **507** |  |
| **454** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **508** |  |
| **455** | different input sources (e.g., RGB versus RAW image). As | | | | | | | | | | | | domain. In this work, xrawt is captured using iPhone Xs | | | | | | | | | | | | | | | | | **509** |  |
| **456** | **510** |  |
| mentioned above, we use the famous YOLOv5 [[16](#page9)] to per- | | | | | | | | | | | | Max, and randomly selected for tests in the discriminator. | | | | | | | | | | | | | | | | |  |
| **457** | **511** |  |
| form object detection. We perform segmentation using the | | | | | | | | | | | | We can optimize the discriminator using cross entropy loss, | | | | | | | | | | | | | | | | |  |
| **458** | **512** |  |
| HRNetv2 [[17](#page9)]. |  |  |  |  |  |  |  |  |  |  |  | given as | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **459** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **513** |  |
| 3.4. Loss Function | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **460** |  |  |  |  |  |  |  |  |  |  |  |  |  |  | Ld = E [log (DRAW (~xrawt ))] | | | | | | | | |  |  |  | (11) | **514** |  |
| **461** | Under the CycleGAN framework, we attempt to simul- | | | | | | | | | | | |  |  |  |  |  |  |  | **515** |  |
|  |  |  |  |  | + E [1 log (DRAW (xrawt ))] | | | | | | | | |  |  | (12) |  |
| **462** |  |  |  |  |  |  |  | **516** |  |
| taneously optimize the restoration quality of both generated | | | | | | | | | | | |  |  |  |  |  |  |  |  |
| **463** |  |  |  |  |  | + E [log (DJPEG (~xjpegs ))] | | | | | | | |  |  |  | (13) | **517** |  |
| RAW image and generated (JPEG coded) RGB image. To- | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |
| **464** |  |  |  |  |  | + E [1 log (DJPEG (xjpegs ))] : | | | | | | | | | |  | (14) | **518** |  |
| wards this goal, we respectively apply the cycle consistency | | | | | | | | | | | |  |  |  |  |  |  |  |
| **465** |  |  |  |  |  |  | **519** |  |
| loss for measuring the cross-domain discrepancy and con- | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **466** | The loss function of YOLOV5 consists of objectness score, | | | | | | | | | | | | | | | | | **520** |  |
| tent loss for context similarity between input source and | | | | | | | | | | | |  |
| **467** | class probability score, and the bounding box regression | | | | | | | | | | | | | | | | | **521** |  |
| generated output. More specifically, we use L1 | | | | | | | | | as cycle | | |  |
| **468** | score. The detection loss are binary cross-entropy with Log- | | | | | | | | | | | | | | | | | **522** |  |
| consistency loss and the well-known VGG [[8](#page9)] based loss | | | | | | | | | | | |  |
| **469** | its loss for class probability and object score. GIOU loss is | | | | | | | | | | | | | | | | | **523** |  |
| (e.g., Lvgg = V GG19 | x0; xxyzg | | | | ) as the content loss. | | | | |  |  |  |
| **470** |  |  | **524** |  |
| In the meantime, | we wish to have the JPEG-converted | | | | | | | | | | | used for the loss calculation of bounding box. | | | | | | | | | | | | | |  |  |  |  |
| **471** |  |  |  |  |  |  |  |  |  |  |  |  | The segmentation loss Lseg is widely used, and could be | | | | | | | | | | | | | | | | **525** |  |
| **472** | RAW output close to the RAW images captured by a target | | | | | | | | | | | |  | **526** |  |
| camera, and the RAW-converted JPEG close to the original | | | | | | | | | | | | formulated as the pixel-wise cross-entropy loss, given as | | | | | | | | | | | | | | | | |  |
| **473** | **527** |  |
| JPEG in native dataset. Thus we introduce an adversarial | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **474** |  |  |  |  |  | H | W | C |  |  |  |  |  |  |  |  |  | **528** |  |
| loss function to minimize the cross-domain discrepancy be- | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **475** |  | Lseg = | | | | X X X | | |  | i |  | | | i |  | | (15) | **529** |  |
| tween source and target images, which can be written as | | | | | | | | | | |  |  |  |  |  | y | log | T |  |  | ; |  |
| **476** |  |  | h=1 w=1 c=1 | | |  | seg; x~upt | | | **530** |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **477** | Ladv = E [1 log (DRAW (~xrawt ))] | | | | | | | |  |  |  | (8) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **531** |  |
| **478** |  |  |  | where H and W denote the height and width of the input | | | | | | | | | | | | | | | | | **532** |  |
| **479** | + E [1 log (DJPEG (~xjpegs ))] : | | | | | | | | |  |  | (9) | image, and C is the number of segmentation classes. | | | | | | | | | | | | | | | |  | **533** |  |
| **480** | It then leads to the total loss function of the generator as | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **534** |  |
| **481** |  | 4. Experiments | | | | | | |  |  |  |  |  |  |  |  |  |  | **535** |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **482** | Lg = L1 + vgg Lvgg + adv Ladv: | | | | | | | |  |  | (10) | |  |  |  |  |  |  |  |  |  |  | **536** |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **483** |  |  |  | We performed comprehensive experiments to demon- | | | | | | | | | | | | | | | | **537** |  |
| **484** | For the discriminator network, we need to distinguish the | | | | | | | | | | | | strate the efficiency of vision tasks execution in the RAW | | | | | | | | | | | | | | | | | **538** |  |
| **485** | x~rawt from the source domain and the xrawt | | | | | | | from the target | | | | | domain. | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **539** |  |



5

ICCV

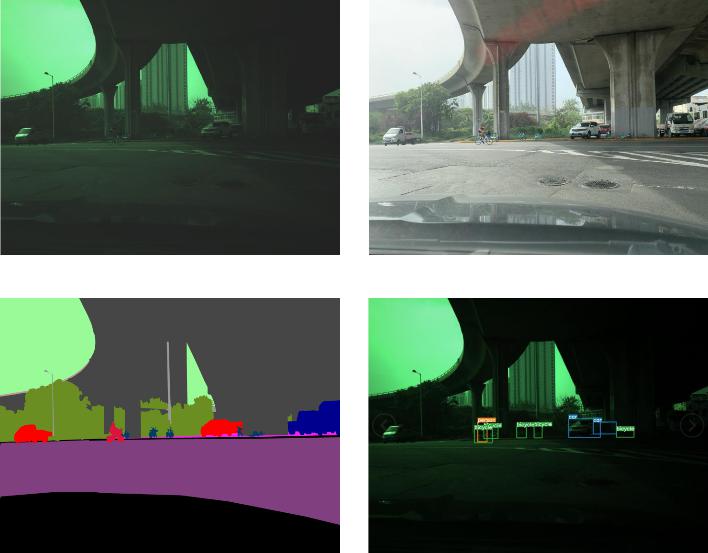
#3769

**ICCV 2021 Submission #3769. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.**

ICCV

#3769

|  |  |  |  |
| --- | --- | --- | --- |
| **540** |  |  |  |
| **541** |  |  |  |
| **542** |  |  |  |
| **543** |  |  |  |
| **544** |  |  |  |
| **545** |  |  |  |
| **546** |  |  |  |
| **547** | (a) | (b) |  |
| **548** |  |
|  |  |  |
| **549** |  |  |  |
| **550** |  |  |  |
| **551** |  |  |  |
| **552** |  |  |  |
| **553** |  |  |  |
| **554** |  |  |  |
| **555** | (c) | (d) |  |
| **556** |  |
| **557** |  |  |  |



1. Figure 4. Examples of iRAW dataset. (a) RAW image. (b) RGB
2. image. (c) Visualized semantic segmentation. (d) Bounding box.

Zoomed in for better details.

**560**

**561**

1. 4.1. Datasets

**563**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Metrics | FID#2 | |  |  | MOS"2 |  |  |
| Dataset | -City | BDD | -City |  |
| BDD | D | D |  |
| RGB | 155.69 | 197.98 | | 0.0 | 0.0 | |  |
| InvGamma [[10](#page9)] | 107.48 | 110.96 | | 1.1 | 1.3 | |  |
| CIE-XYZ Net [[2](#page9)] | 90.68 | 133.86 | | 2 | 2.6 | |  |
| CycleISP [[21](#page9)] | 93.09 | 106.28 | | 3.6 | 3.2 | |  |
| Our invISP | 83.37 | 92.56 | | 4 | 4.1 | |  |

Table 1. The FID & MOS results of InvGamma [[10](#page9)], CIE-XYZ

Net [[2](#page9)], CycleISP [[21](#page9)] and our GAN-based invISP on BDD and D2-City.

large-scale auto driving video dataset, providing more than 10,000 forward-looking video data recorded by the dash-cam in China cities. All videos are recorded in either 720p or 1080p resolution at 30 FPS. Note that D2-City provides frame by frame annotations for nearly 1000 videos, includ-ing target frame location, target category and tracking ID information, covering a total of 12 categories. Since D2-City did not provide the key frames of videos, we use inter frame similarity algorithm to extract the 24,706 key frames. Finally, we combine BDD 100k and D2-City as our training dataset.

**594**

**595**

**596**

**597**

**598**

**599**

**600**

**601**

**602**

**603**

**604**

**605**

**606**

**607**

**608**

**609**

**610**

**611**

**612**

**613**

**614**

**615**

**616**

**617**

1. iPhone RAW Scape 1K. To study the RAW-domain vi-
2. sion tasks, we created a RAW-domain verification dataset,
3. iPhone RAW Scape 1K (iRAW). The entire dataset was shot
4. using the iPhone XS Max mobile phone fixed on the dash-
5. board in a car to simulate the autonomous driving scenario.
6. We collected 1153 RAW images at 4k 3k resolution and
7. generated corresponding RGB images using built-in ISP.
8. We manually labeled all detection bounding boxes, includ-
9. ing car, person, bicycle, and motorcycle. We marked all the
10. images with panoramic segmentation including sixteen cat-
11. egories of road, sidewalk, building, wall, fence, pole, traf-
12. fic light, traffic sign, plant, sky, person, rider, car, truck,
13. bus, and cylce (which includes bicycle, motorcycle, tricy-
14. cle). Figure [4](#page6) shows an example image from our dataset.
15. For object detection, we randomly selected 10% RAW im-
16. ages as the test set and used the remaining to fine-tune the
17. YOLOv5 model. Since the amount of data is sufficient for
18. full-scene segmentation, we train and test directly on iRAW.
19. BDD & D2-City. We collected nearly 1K RAW images
20. in iRAW, but they are insufficient to train a reliable detec-
21. tion network model. Therefore, we propose to convert the
22. existing large-scale RGB datasets such as BDD 100K [[20](#page9)],
23. which contains 100,000 720p30FPS (frame per second)
24. videos, into RAW-domain representation for pre-training. It
25. covers the image data captured from a variety of US major
26. metro areas including the New York City, San Francisco,
27. and Bay area. It also includes various scenes such as the
28. city street, residential area, and highway. In total, we re-
29. trieve 79,326 key frames provided by BDD as training im-
30. ages. Additionally, we use the D2-City [[5](#page9)], which is another

4.2. Training Details

Learned invISP Network. Because of the significant differences between BDD and D2-City, we train two sep-arate GANs for the two datasets respectively. The lin-ear transformation network and the nonlinear transforma-tion network were pre-trained for 300 epochs on the iRAW dataset. Meanwhile, w We randomly generated the JPEG compressed RGB images on the iRAW dataset using the quality factor Q 2 [1; 100], and the JPEG artifacts removal network was pre-trained for 200 epochs on the paired data. Adam optimizer with batch size 2 is applied for training, and the initial learning rate is set to 0.00005 and 0.0001 for the generator and discriminator, respectively. We set the to-tal number of iteration to 100k. vgg is set to 0.01, and adv is set to 10.

Object Detection Network. We deploy YOLOv5 [[6](#page9)] as the baseline model. The input image is resized to 1024 704 for training. After training 100 epochs on BDD and D2-City, we fine-tune with iRAW dataset for another 20 epochs.

Segmentation Network. We use HRNetv2 [[18](#page9)] as our baseline model. The input image is resized to 1024 512. For rearranged xrerawg , its input size is 512 256 4. We employ SGD optimizer with batch size 24 and the initial learning rate at 0.0025. Following the configuration in [[18](#page9)], we decay learning rate at every iteration after first 150k it-erations. For the segmentation network, we only train our models on iRAW dataset.

**618**

**619**

**620**

**621**

**622**

**623**

**624**

**625**

**626**

**627**

**628**

**629**

**630**

**631**

**632**

**633**

**634**

**635**

**636**

**637**

**638**

**639**

**640**

**641**

**642**

**643**

**644**

**645**

**646**

**647**

6

ICCV ICCV

#3769 #3769

**ICCV 2021 Submission #3769. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.**

**648**

**649**

**650**

**651**

**652**

**653**

**654**

**655**

**656**

**657**

**658**

**659**

**660**

**661**

**662**

**663**

**664**

**665**

**666**

**667**

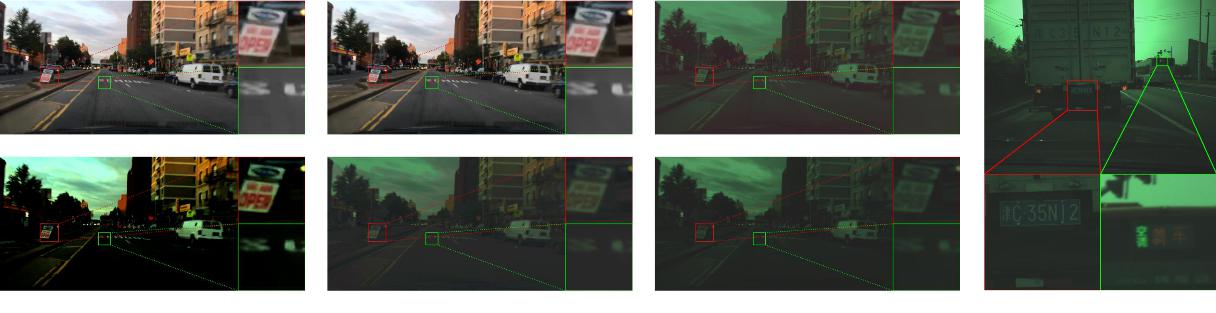
|  |  |  |
| --- | --- | --- |
| BDD | JPEG |  |
|  |  |

InvGamma

|  |  |
| --- | --- |
| D 2 -City | JPEG |

InvGamma

JPEG Artifacts Removal our GAN-based invISP



CIE-XYZ Net CycleISP Real RAW



|  |  |
| --- | --- |
| JPEG Artifacts Removal | our GAN-based invISP |

CIE-XYZ Net CycleISP Real RAW

**702**

**703**

**704**

**705**

**706**

**707**

**708**

**709**

**710**

**711**

**712**

**713**

**714**

**715**

**716**

**717**

**718**

**719**

**720**

**721**

1. Figure 5. Visualization of RAW images generated by InvGamma [[10](#page9)], CIE-XYZ Net [[2](#page9)], CycleISP [[21](#page9)] and our GAN-based invISP on
2. BDD and D2-City. Real RAW images from iRAW dataset are presented for comparison.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **670** |  |  |  |  |  |  |  |  |  |  |  |
| **671** | Domain | Pretrain | Finetuning | Test | car | person | bicycle | motorcycle | delay | mAP0.5 |  |
| **672** | RGB | BDD + D2-City | iRAWtrainRGB set | iRAWtestRGB set | 90.9 | 80.3 | 67.0 | 76.1 | 35.7ms | 78.6 |  |
| **673** | RAW | RAWinvISP | iRAWtrain set | iRAWtest set | 90.6 | 79.2 | 69.0 | 69.4 | 2.4ms | 77.1 |  |
|  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **674** | Table 2. Comparative studies of object detection in RGB domain and RAW domain. | | |  |
| **675** |  |  |  |  |
| **676** | 4.3. Results | 4.3.2 | Detection Results |  |
| **677** |  |
| **678** | 4.3.1 Learned invISP | As shown in Table [2](#page7) and Figure [6](#page8), we see that the ISP delay | |  |
| **679** |  |
|  | is significantly reduced. Using a 30 FPS camera as an ex- | |  |
| **680** | Figure [5](#page7) presents the results of converting the RGB im- |  |
| ample, executing CV tasks in RAW domain can save 93% | |  |
| **681** | ages in BDD & D2-City dataset into RAW images us- |  |
| time compared to processing in the RGB domain with little | |  |
| **682** | ing three methods: our proposed GAN-based invISP, in- |  |
| loss in performance (-1.5 mAP). In some categories, such | |  |
| **683** | vGamma [[10](#page9)], CIE-XYZ Net [[2](#page9)], and CycleISP [[21](#page9)]. To |  |
| as cycle, the detection accuracy in RAW domain exceeds | |  |
| **684** | be fair, all compared networks were finetune on our iRAW |  |
| accuracy in the RGB domain. | |  |
| **685** | dataset with 300 epochs. We observe that the color and de- |  |
|  |  |  |
| **686** | tails of the learned invISP generated images are closer to | 4.3.3 | Segmentation Results |  |
| **687** | the RAW images from iRAW dataset. This suggests the ef- |  |
| **688** | ficiency of GAN-based invISP model that can clearly learn | Table [3](#page8) presents segmentation results in the RAW and RGB | |  |
| **689** | the cross-domain information and remove the JPEG arti- |  |
| domains for the six categories: road, traffic light, traffic | |  |
| **690** | fact. To evaluate the quality of our invISP generated images, |  |
| sign, person, rider and car. The results show that the seg- | |  |
| **691** | we use Frechet´ inception distance (FID) and Mean Opinion |  |
| mentation task in the RAW domain reduces latency with lit- | |  |
| **692** | Score (MOS). FID is commonly used in GAN to measure |  |
| tle loss (-1.3 mIOU) in segmentation performance. Figure [6](#page8) | |  |
| **693** | the distribution gap between the generated images and the |  |
| shows example results of our segmentation task in iRAW. | |  |
| **694** | source domain images. Table [1](#page6) shows that our method pro- |  |
| We can see that the segmentation results of iRAW is very | |  |
| **695** | vides FID score nearly 10 points lower than that of Cycle- |  |
| close to that of the RGB segmentation. | |  |
|  |  |  |

1. GAN. For MOS, we invited 5 testers to score our generated

|  |  |  |  |
| --- | --- | --- | --- |
| **697** | images, with a score of 5 if they cannot distinguish whether | 5. Conclusion and Discussion |  |
|  |  |  |

1. they are RAW images captured by iPhone XS Max or 0 if

|  |  |  |
| --- | --- | --- |
| **699** | they are closer to the original JPEG images. Results sug- | In this paper, we explore and verify the feasibility of ex- |
| **700** | gest that our images generated by our proposed invISP had | ecuting high-level computer vision tasks in RAW domain. |
| **701** | better subjective similarity to RAW images. | In order to verify the potential of RAW processing in real- |

**722**

**723**

**724**

**725**

**726**

**727**

**728**

**729**

**730**

**731**

**732**

**733**

**734**

**735**

**736**

**737**

**738**

**739**

**740**

**741**

**742**

**743**

**744**

**745**

**746**

**747**

**748**

**749**

**750**

**751**

**752**

**753**

**754**

**755**

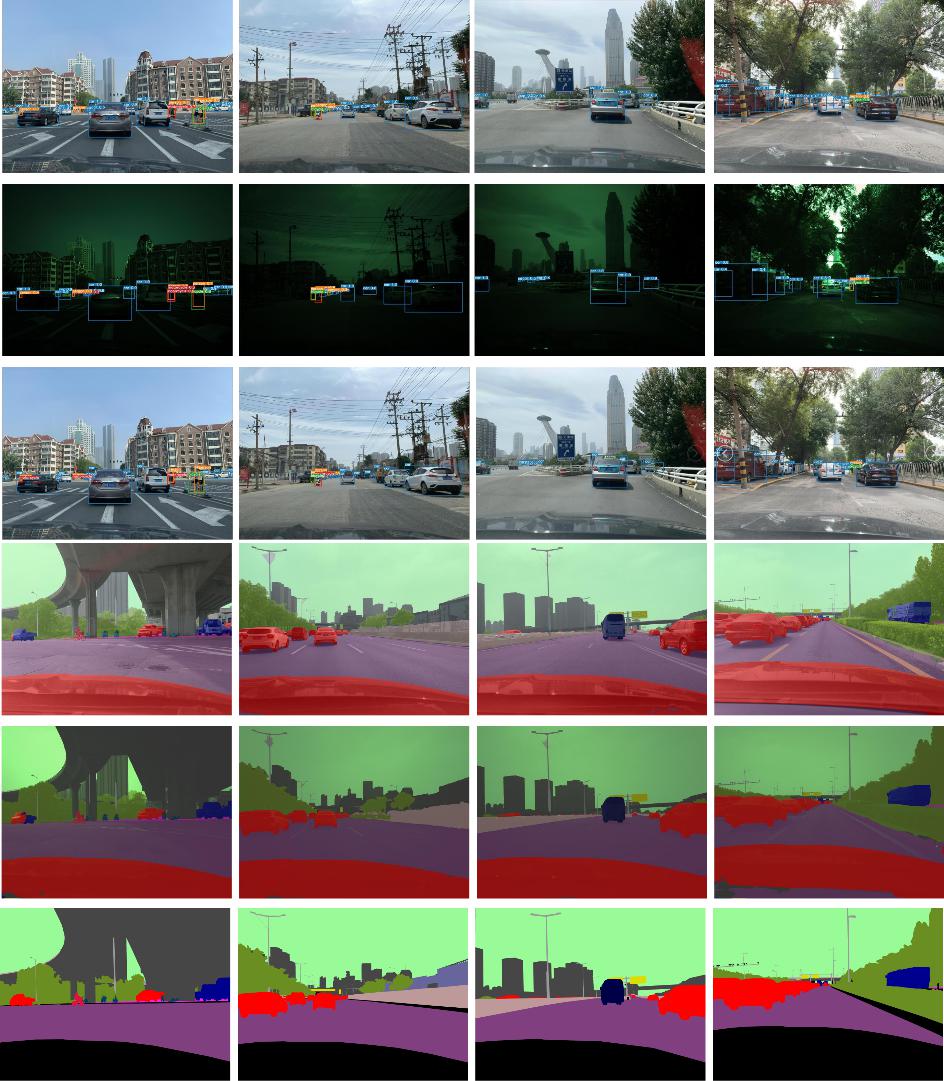
7

ICCV ICCV

#3769 #3769

**ICCV 2021 Submission #3769. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **756** |  | Domain | Road | Sidewalk |  | Building | Wall | Fence | Pole | Traffic | Traffic | Plant | Sky | Person | Rider | Car | Truck | Bus | Cycle | Delay | mIoU |  |
| **757** |  |  | Light | Sign |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **758** |  | RGB | 97.29 | 60.57 |  | 91.34 | 40.88 | 84.57 | 69.57 | 54.80 | 86.55 | 94.57 | 99.02 | 59.56 | 64.43 | 92.28 | 84.02 | 82.24 | 46.91 | 955.7ms | 75.54 |  |
|  | RAW | 97.48 | 60.31 |  | 90.63 | 36.58 | 84.17 | 68.4 | 57.69 | 87.55 | 94.07 | 98.99 | 59.85 | 66.29 | 93.1 | 70.72 | 75.77 | 46.19 | 922.4ms | 74.24 |  |
| **759** |  |  |  |
|  |  |  |  | Table 3. Comparative studies of semantic segmentation in RGB domain and RAW domain. | | | | | | | | | | | | | |  |  |  |  |
| **760** |  |  |  |  |  |  |  |  |
| **761** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **762** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **763** |  |  |  | Detection | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **764** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  | in RGB | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **765** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **766** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **767** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **768** |  |  |  | Detection | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **769** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  | in RAW | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **770** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **771** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **772** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **773** |  |  |  | Ground | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **774** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  | Truth | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **775** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **776** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **777** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **778** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **779** |  |  |  | Segmentation | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **780** |  |  |  | in RGB | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **781** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **782** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **783** |  |  |  | Segmentation | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **784** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  | in RAW | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **785** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **786** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **787** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **788** |  |  |  | Ground | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **789** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  | Truth | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **790** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **791** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **792** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **793** |  |  |  | Figure 6. The visualization of object detection and semantic segmentation in both RGB and RAW domain. | | | | | | | | | | | | | | | |  |  |  |
| **794** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **795** | world, we have collected and labeled the first RAW domain | | | | | | | | | | | rectly on it. Therefore, we hope to collect a larger RAW | | | | | | | | | |  |
| **796** |  |
| dataset iRAW. Due to the fact that most datasets are RGB | | | | | | | | | | | dataset for direct training to better evaluate the gap between | | | | | | | | | |  |
| **797** |  |
| format, we propose a GAN based learnt inverse ISP to ex- | | | | | | | | | | | CV tasks in RAW and RGB domain. | | | | | | |  |  |  |  |
| **798** |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |



1. tend the CV tasks in RAW domain. Experimental results

|  |  |  |  |
| --- | --- | --- | --- |
| **800** | show that CV tasks in RAW domain can effectively reduce | At the same time, our network architecture completely |  |
| the time delay and calculation, thus potentially saving the |  |
| **801** | uses the network architecture that performs tasks in the |  |
| power with less computation. |  |
| **802** | RGB domain. However, because the characteristics of |  |
|  |  |
| **803** | The performance gap is very small, e.g., 1%, for the ac- | RAW domain are different from that of RGB domain, |  |
| **804** | curacy of CV tasks in RAW domain and that in RGB do- | whether there is a network architecture more suitable for |  |
| **805** | main. we think this is because although we use GAN to | RAW domain is also a question worthy of discussion. At |  |
| **806** | narrow the gap between inverse RAW and real RAW, it can | the same time, for tasks that depend on the input resolution, |  |
| **807** | not be completely eliminated. This can be solved by train- | where to add upsample in the network to restore the input |  |
| **808** | ing directly on real RAW, but due to the cost of acquisition, | resolution and reduce the amount of calculation is also our |  |
| **809** | the number of iRAW datasets is still not enough to train di- | next research direction. |  |

**810**

**811**

**812**

**813**

**814**

**815**

**816**

**817**

**818**

**819**

**820**

**821**

**822**

**823**

**824**

**825**

**826**

**827**

**828**

**829**

**830**

**831**

**832**

**833**

**834**

**835**

**836**

**837**

**838**

**839**

**840**

**841**

**842**

**843**

**844**

**845**

**846**

**847**

**848**

**849**

**850**

**851**

**852**

**853**

**854**

**855**

**856**

**857**

**858**

**859**

**860**

**861**

**862**

**863**

8

ICCV

#3769

**ICCV 2021 Submission #3769. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.**

ICCV

#3769

1. References

**865**

1. [1] Abdelrahman Abdelhamed, Mahmoud Afifi, Radu Timofte,
2. and Michael S Brown. Ntire 2020 challenge on real image
3. denoising: Dataset, methods and results. In Proceedings of
4. the IEEE/CVF Conference on Computer Vision and Pattern

Recognition Workshops, pages 496–497, 2020. [2](#page2)

**870**

[2] Mahmoud Afifi, Abdelrahman Abdelhamed, Abdullah

**871**

Abuolaim, Abhijith Punnappurath, and Michael S Brown.

1. Cie xyz net: Unprocessing images for low-level computer
2. vision tasks. arXiv preprint arXiv:2006.12709, 2020. [2](#page2), [3](#page3),
3. [4](#page4), [6](#page6), [7](#page7)
4. [3] David J Brady, Lu Fang, and Zhan Ma. Deep learning for
5. camera data aquisition, control and image estimation. Ad-
6. vances in Optics and Photonics, Sept. 2020. [1](#page1)
7. [4] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen,
8. Dillon Sharlet, and Jonathan T Barron. Unprocessing images
9. for learned raw denoising. In Proceedings of the IEEE Con-
10. ference on Computer Vision and Pattern Recognition, pages
11. 11036–11045, 2019. [2](#page2)
12. [5] Zhengping Che, Guangyu Li, Tracy Li, Bo Jiang, Xuefeng Shi, Xinsheng Zhang, Ying Lu, Guobin Wu, Yan Liu, and
13. Jieping Ye. D2-city: A large-scale dashcam video dataset of

**885**

diverse traffic scenarios. arXiv preprint arXiv:1904.01975,

1. 2019. [6](#page6)
2. [6] Jocher Glenn, Stoken Alex, Borovec Jirka, NanoCode012,
3. ChristopherSTAN, Changyu Liu, Laughing, tkianai, Hogan
4. Adam, lorenzomammana, yxNONG, AlexWang1900, Di-
5. aconu Laurentiu, Marc, wanghaoyang0106, ml5ah, Doug,
6. Ingham Francisco, Frederik, Guilhen, Hatovix, Poznanski
7. Jake, Fang Jiacong, Yu Lijun, changyu98, Wang Mingyu,
8. Gupta Naman, Akhtar Osama, PetrDvoracek, and Rai
9. Prashant. ultralytics/yolov5: v3.1 - Bug Fixes and Perfor-

mance Improvements, Oct. 2020. [6](#page6)

**895**

[7] Guoheng Huang, Junwen Zhu, Jiajian Li, Zhuowei Wang,

**896**

Lianglun Cheng, Lizhi Liu, Haojiang Li, and Jian Zhou.

**897**

Channel-attention u-net: Channel attention mechanism for

1. semantic segmentation of esophagus and esophageal cancer.
2. IEEE Access, 8:122798–122810, 2020. [4](#page4)
3. [8] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual
4. losses for real-time style transfer and super-resolution. In
5. European conference on computer vision, pages 694–711.
6. Springer, 2016. [5](#page5)
7. [9] Hakki Can Karaimer and Michael S Brown. A software
8. platform for manipulating the camera imaging pipeline. In
9. European Conference on Computer Vision, pages 429–444.
10. Springer, 2016. [2](#page2)
11. [10] Samu Koskinen, Dan Yang, and Joni-Kristian Kam¨ar¨ainen¨.

Reverse imaging pipeline for raw rgb image augmentation.

**909**

In 2019 IEEE International Conference on Image Processing

**910**

(ICIP), pages 2896–2900. IEEE, 2019. [2](#page2), [6](#page6), [7](#page7)

**911**

[11] Jianwei Li, Yongtao Wang, Haihua Xie, and Kai-Kuang Ma.

1. Learning a single model with a wide range of quality fac-
2. tors for jpeg image artifacts removal. IEEE Transactions on
3. Image Processing, 29:8842–8854, 2020. [4](#page4)
4. [12] Zhihao Li and Zhan Ma. Robust white balance estimation
5. using joint attention and angular loss optimization. In Thir-
6. teenth International Conference on Machine Vision, volume

11605, page 116051E. International Society for Optics and Photonics, 2021. [4](#page4)

1. Junichi Nakamura. Image sensors and signal processing for digital still cameras. CRC press, 2017. [1](#page1)
2. Rang MH Nguyen and Michael S Brown. Raw image recon-struction using a self-contained srgb-jpeg image with only 64 kb overhead. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1655– 1663, 2016. [2](#page2)
3. S.J.D. Prince. Computer Vision: Models Learning and Infer-ence. Cambridge University Press, 2012. [1](#page1)
4. Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object de-tection. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 779–788, 2016. [2](#page2), [5](#page5)
5. Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose esti-mation. In CVPR, 2019. [2](#page2), [5](#page5)
6. Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514, 2019. [6](#page6)
7. Sifeng Xia, Kunchangtai Liang, Wenhan Yang, Ling-Yu Duan, and Jiaying Liu. An emerging coding paradigm vcm: A scalable coding approach beyond feature and signal. In 2020 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6. IEEE, 2020. [1](#page1)
8. Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Dar-rell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF Con-ference on Computer Vision and Pattern Recognition, pages 2636–2645, 2020. [6](#page6)
9. Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2696– 2705, 2020. [2](#page2), [3](#page3), [6](#page6), [7](#page7)
10. Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augus-tus Odena. Self-attention generative adversarial networks. In International conference on machine learning, pages 7354– 7363. PMLR, 2019. [4](#page4)
11. Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 2223– 2232, 2017. [2](#page2), [3](#page3), [4](#page4)

**918**

**919**

**920**

**921**

**922**

**923**

**924**

**925**

**926**

**927**

**928**

**929**

**930**

**931**

**932**

**933**

**934**

**935**

**936**

**937**

**938**

**939**

**940**

**941**

**942**

**943**

**944**

**945**

**946**

**947**

**948**

**949**

**950**

**951**

**952**

**953**

**954**

**955**

**956**

**957**

**958**

**959**

**960**

**961**

**962**

**963**

**964**

**965**

**966**

**967**

**968**

**969**

**970**

**971**

9