RANUS: RGB and NIR Urban Scene Dataset for Deep Scene Parsing

Gyeongmin Choe[1], Seong-Heum Kim[1], Sunghoon Im[1], Joon-Young Lee[2], Srinivasa G. Narasimhan[3] and In So Kweon[1]

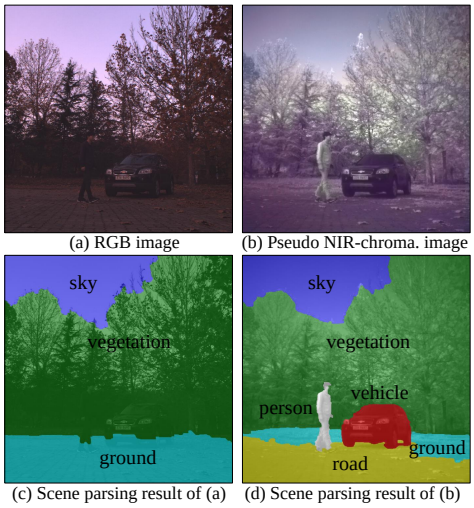
***Abstract*—In this paper, we present a data-driven method for scene parsing of road scenes to utilize single-channel near-infrared (NIR) images. To overcome the lack of data problem in non-RGB spectrum, we define a new color space and decompose the task of deep scene parsing into two sub-tasks with two separate CNN architectures for chromaticity channels and semantic masks. For chromaticity estimation, we build a spatially-aligned RGB-NIR image database (40k urban scenes) to infer color information from RGB-NIR spectrum learning process and leverage existing scene parsing networks trained over already available RGB masks. From our database, we sample key frames and manually annotate them (4k ground truth masks) to finetune the network into the proposed color space. Hence, the key contribution of this work is to replace multispectral scene parsing methods with a simple yet effective approach using single NIR images. The benefits of using our algorithm and dataset are confirmed in the qualitative and quantitative experiments.**

***Index Terms*—Deep Learning in Robotics and Automation, Semantic Scene Understanding**

I. Introduction

RECENT deep learning frameworks for scene parsing have shown significant advances in RGB images. That said, several challenging situations remain where current methods often fail; *e.g.*low-light situations such as those at dusk or dawn; objects exhibiting poor contrast against a background; and poor-visibility conditions such as haze, mist, and fog. See examples in Figure 1 and Figure 2. These failures are still too common for safety critical applications such as autonomous navigation of vehicles. Meanwhile, the near infrared (NIR) images capture the unique value of chlorophyll in natural background and show better contrast in the presence of sky or haze, as discussed in earlier work [1]. We exploit the benefits of images captured in NIR spectrum for road scene parsing.

Although there are several advantages of using an NIR camera in the challenging situations described above, one particular problem when we extending the NIR imaging to a deep learning framework for vehicle applications is lack of proper database in non-RGB spectrum. To the best of our knowledge, it is the first work to use the data-driven method in segmenting NIR images. To overcome the lack of data issue, we create a new color space and decompose the task of deep scene parsing into two sub-tasks with two separate CNN architectures for chromaticity channels and semantic masks. Compared to the work [1], our deep learning-based scene parsing framework needs only a single-channel NIR camera and replaces color information with automatically-generated RGB chromaticity cue.



(a) RGB image (b) Pseudo NIR-chroma. image (c) Scene parsing result of (a) (d) Scene parsing result of (b)

Fig. 1. Result comparison of (a): normal RGB image and (b): our pseudo NIR-chromaticity image. Note that (b) is our colorized result from single-channel NIR image. In (a), a black car, a pedestrian wearing dark clothing and background vegetations are ambiguously seen, which yield incorrect scene parsing in (c). Whereas, in (d), pedestrian and car are accurately detected and clearly segmented.

In the first phase, we begin by building an RGB-NIR video database of real-world road scenes where corresponding RGB and NIR frames are spatially aligned. Our database contains 40k scenes captured from a vehicle driven across various urban, rural and campus roads consisting of buildings and natural landscapes. Using this database, we generate NIR-chromaticity images by augmenting the NIR brightness with the chromaticity of RGB images. In order to estimate two chromaticity channels from RGB-NIR spectrum learning process, we train a network that maps the NIR brightness to the RGB chromaticity. We use a baseline colorization network [3] that infers visible spectrum color channels from a grayscale image. We fine-tune the model with our RGB-NIR

Manuscript received: August 18, 2017; Revised December 15, 2017; Accepted January 15, 2018.

This paper was recommended for publication by Editor Jana Kosecka upon evaluation of the Associate Editor and Reviewers' comments.

This research was supported by the Ministry of Trade, Industry & Energy and the Korea Evaluation Institute of Industrial Technology (KEIT) with the program number of 10060110.

[1]G.Choe, S-H.Kim, S.Im and I.S.Kweon are with the School of Electrical Engineering, KAIST, Daejeon, Republic of Korea. {gmchoe08, seongheum.kim}@gmail.com; iskweon@kaist.ac.kr (corresponding author: S.-H. Kim)

[2]J-Y. Lee is with the Adobe Research, San Jose, CA, USA jolee@adobe.com

[3]S.G. Narasimhan is with the Carnegie Mellon University, PA, USA srinivas@cs.cmu.edu

Digital Object Identifier (DOI): see top of this page.

Fig. 2. Benefits using NIR camera in challenging real-world scenes. (a-e): From our database. (f): From [2].

database and obtain pseudo NIR-chromaticity images from single-channel NIR images.

After that, the pseudo NIR-chromaticity images are fed into a scene parsing network, such as the DeepLab-CRF method [4] which uses a residual-net pre-trained on MS-COCO [5]. We train the network with additional databases such as CamVid [6], PASCAL-context [7], and Cityscapes [8]. From our database, we also sampled 4k key frames and manually labeled them at the pixel-level to provide ground truth segmentation masks. These ground truth masks are used to fine-tune the RGB-trained network so that we can alleviate the brightness difference between the visible and NIR band. We evaluate the scene parsing performances on challenging scenes, analyze key factors of our algorithm and demonstrate that our pseudo NIR-chromaticity image provides better scene parsing results than solely using RGB or NIR image.

The major contributions of our work are outlined in the followings. First, we build RGB and NIR urban scene dataset (RANUS) for vehicle applications, consisting of 40k precisely aligned scenes and 4k manually annotated masks. Second, we develop a deep scene parsing of non-RGB images by utilizing spectrum learning process and fine-tuning the network into the target domain. Last, we validate the effectiveness of using single NIR images in improving deep semantic segmentation of the scenes, which are often challenging to RGB images.

II. Related Work

NIR imaging for Vision: NIR imaging has been increasingly used for various computer vision tasks. Works in [9], [10] jointly capture NIR and visible images. Compared to visible images, NIR images have various advantages and are used for shadow detection [11], scene category classification [2], depth enhancement [12], and reflectance estimation [13]. In [14], NIR images are used for the haze removal of RGB image. They merge the haze-free NIR intensity with corresponding RGB image to obtain a clear image. In [2], Brown and Süsstrunk capture RGB and NIR image pairs by attaching different filters in turn. This database includes images captured from outdoors, but contains only static scenes with limited number (400). It should also be noted that NIR imaging has been increasingly used for 3D sensing (*e.g.*Kinect, RealSense) and has been used jointly with RGB data for better indoor scene parsing. Our focus in this work remains outdoor road scene understanding where the range and accuracy of such 3D sensors are severely limited.

Image Colorization: Inferring color from gray images has started from the affinity-based optimization method [15]. It shows good results but requires users to scribble colors in an image. Nowadays, exploiting deep features has become an alternative way to infer pixel-level color information [3], [16]. These works commonly use large scale data and train a model with a convolutional neural network framework, however they have differences on loss functions and data usages. The method in [16] is based on the VGG network [17] which is reinforced with hypercolumns. The work in [3] uses a similar network architecture, but poses the colorization as a classification problem with a rebalanced classification loss. In [18], [19], RGB and NIR images are captured in turn. They show the visual enhancement of the RGB image using NIR image, but does not infer color automatically, and targets only static scenes. Concurrent work [20] also tries to colorize an infrared image but shows strong blur artifacts.

Scene Parsing: Most research on semantic segmentation now adopt deep learning based techniques [21], [22] as rapidly replacing hand-craft feature based methods [23], [24]. Farabet *et al.*[25] apply DCNNs to multiple resolutions of the image. Chen *et al.* [26] combine the CNN with a fully connected conditional random field [24] to raise the accuracy of the segmentation. Eigen and Fergus [27] refine the prediction result with coarse-to-fine scheme and propagate the coarse result. More recently, in [4], Chen *et al.*use ResNet [28] as a base network model, which performs better than [26] based on VGG-16 [17]. In [1], [29], Salamati *et al.*utilize hand craft features of RGB and NIR jointly and demonstrate the benefits of NIR. Motivated from the work, we extend the NIR imaging in the work [1] to a recent deep learning framework.

III. Revisiting Useful Properties of NIR Imaging

In this section, we first present salient differences between RGB and NIR scene appearances and revisit the benefits of NIR imaging for vehicle applications.

A. High Contrast of Natural and Artificial Objects

In the NIR band, man-made objects and constructions (*e.g.*vehicles, buildings, roads) show mostly consistent lightness with that of the visible band, whereas vegetations (*e.g.*tree, grass, mountain) are far brighter than in an RGB image [30], [1]. NIR images often capture higher contrast of foreground objects against background as compared to RGB images. For example, people wearing dark clothing appear much brighter in NIR images [9], [13]. In Figure 2 (a-d), we depict this phenomenon in real-world road scenes. A traffic speed camera in (a), the tunnel in (b) and the car in (c) are clearly detected in the NIR image, whereas they are incorrectly

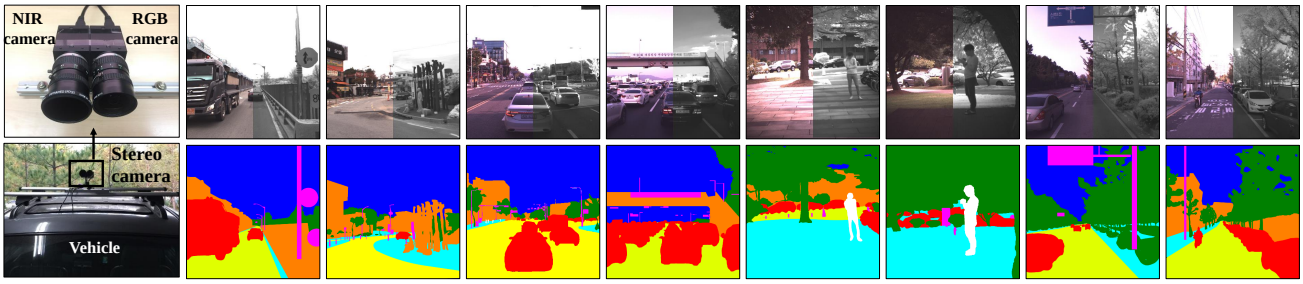


Fig. 3. Our RANUS database contains 50 videos, 40k images for each RGB and NIR.. The entire RGB and NIR pairs are spatially aligned. Among the images, we provide ground-truth semantic segmentation masks for 4k key frames, which are manually annotated. We visualize some of the example images; In first row, we visualize images whose left half shows the RGB and the right half shows the NIR image. Second row shows the corresponding ground truth scene parsing label images.

labeled as background vegetations in the RGB image. Also in (d), we see pedestrians wearing dark clothing clearly detected against the construction in the NIR image.

B. Less Atmospheric Scattering

Based on Rayleigh scattering [31], NIR light tends to undergo less atmospheric scattering under poor weather conditions such as haze or through in other types of particulate matter (pollutants, dust, underwater impurities) [14]. This phenomenon is also observed in our experiments. In Figure 2 (e-f), atmospheric haze or particulate matter is less captured in NIR image. We revisit these useful properties of NIR imaging for vehicle applications.

IV. RGB AND NIR URBAN SCENE DATASET (RANUS)

Our database is named as RANUS (RGB and NIR urban scene stereo) dataset. This section describes our capture setup, the dataset captured and a method to align RGB-NIR image pairs and a method to annotate ground-truth segmentation masks.

A. Data Acquisition

Our capture setup consists of a stereo rig with one RGB- and one NIR- sensitive camera. Note that CMOS/CCD sensors are also sensitive to the NIR spectrum and that the same sensor is often used with a NIR block filter or a NIR pass filter. It is thus possible to match the type of imaging sensor, the camera size, the sensor resolution, and the mounted lenses for both cameras. This simplifies the calibration process and is beneficial for seamless alignment. In our work, we use two Point-grey grasshopper cameras (NIR: GS3-U3-41C6NIR-C, RGB: GS3-U3-41C6C-C). Those two cameras are horizontally separated and mounted on the roof of a vehicle, as shown in Figure 3. While the vehicle is driving through various places covering urban, rural, campus roads and nature scenes consist of buildings and mountains, the cameras simultaneously capture the scenes at a rate of 10fps. In total, 50 videos with 40,000 frames were collected.

B. Image Alignment

After capturing the data, we spatially align the RGB and NIR pair images. The baseline of our stereo setup is much shorter than the depth of the road scene, which yields less occlusions. We found that a simple homography-based warping method cannot precisely align the stereo pairs over the entire field of view because the captured scenes contain a wide depth range (see the supplementary material). Therefore, we use a pixel-wise image alignment process based on the flow field, which is more suitable for our RGB-NIR stereo setup. As an initial step, we rectify the RGB-NIR pairs with pre-calibrated camera parameters to constrain the correspondence search along horizontal epipolar lines (image rows). Moreover, we can set the x-axial search range $[0, d_{max}]$ in terms of the baseline between the two cameras B and the minimum depth in the scene Z_{min}:

$$d_{max} = Bf/Z_{min}, \tag{1}$$

where, f is focal length of the camera, and Z_{min} is set as the distance from the camera to the front-end of the vehicle. With the rectified images, we construct dense descriptors using Dense Adaptive Self-Correlation (DASC) [32], primarily designed for multi-modal or multi-spectral image pairs. Subsequently, we exploit the SIFT flow matching scheme [33] with the specified search range of our camera setup. Using this approach, we obtain two flow maps based on the reversible match hypothesis and assess the left and right consistency to reinforce unreliable pixels with a tree-based propagation approach [34]. Several examples of aligned RGB-NIR images in RANUS are shown in Figure 3, covering a wide range of road scenes. In each image, the left half shows the RGB image and right half shows the corresponding aligned NIR image.

C. Manual Annotation

Among the entire images, we sample 4000 key frames and manually annotate semantic segmentation masks at the pixel-level to provide ground-truth data for NIR-chromaticity images. We use these ground-truth masks for training the scene parsing networks and evaluating their performances. For the manual labeling process, pixel-level labels were annotated by 30 students using the pen tool in Photoshop for 4k images. Our approach also takes advantage of previously annotated RGB data from CamVid [6], PASCAL-context [7], and Cityscapes [8]; each dataset has different types of labels. Hence, we take the common classes of the datasets and redefine the class labels into the ten classes of SKY, GROUND, WATER, MOUNTAIN, ROAD, CONSTRUCTION, VEGETATION, OBJECT, VEHICLE, and PEDESTRIAN, which are suitable for our driving scenes. Several examples are shown in Figure 3.

V. NIR-CHROMATICITY FOR DEEP SCENE PARSING

In this section, we propose the use of *NIR-chromaticity* images for scene parsing tasks.

A. NIR-chromaticity Images

In the Lab color space, an RGB image is re-structured with lightness l_c, and two color components a_c and b_c. The NIR-chromaticity image I_s augments color from the RGB images a_c and b_c with lightness from the NIR image l_n , which yields $I_s = l_n a_c b_c$. For challenging scenes such as low-light, ambiguous contrasts, and poor visibility conditions in the RGB image, we exploit the NIR advantages by replacing the luminance. In order to validate this idea, we utilize a state-of-the-art CNN architecture for semantic image segmentation. It is already well known that deep learning can gain benefits from a large amount of annotated data. Our problem is no exception, but only a few segmentation masks are available for multi-band datasets (*e.g.*RGB+NIR). Hence, our approach of using NIR-chromaticity image (3-channel) can be used together with the massive RGB data instead of using 4-channel inputs to train network from scratch.

B. Deep Scene Parsing

Our scene parsing approach is based on DeepLab-CRF pipeline [4] which estimates the label distribution per pixel using VGG16 or a residual network (ResNet) and post-processes the label predictions with a fully connected Conditional Random Field (CRF). The CRF parameters are learnt by cross-validation to maximize the intersection-over-union (IOU) segmentation accuracy in fully-annotated images. Similar to original work, the dense CRF in this implementation is not trained in an end-to-end manner and is decoupled from the DCNN in the training stage. Extending the refinement process in the testing phase, the CRF inferences are considered as the initial regions of the appearance models for each label, followed by standard graph optimization of the per-pixel distributions [35]. For this additional post-processing, we consider the NIR part of the spectrum to the local color model with an instance region that learns the specific appearance of object boundaries (if the NIR channel exists). Given the multi-spectral data and the initial label prediction from the network, we iteratively polish the object boundaries and update the estimated regions, as shown in Figure 4.

The scene parsing network is trained over MS-COCO [5], CamVid, PASCAL-context, and Cityscapes datasets in accordance with our re-definition of the classes, which are SKY, GROUND, WATER, MOUNTAIN, ROAD, CONSTRUCTION, VEGETATION, OBJECT, VEHICLE, and PEDESTRIAN. In addition to the fact that our approach can capture unique patterns and contexts of the classes through CNN features, our training data is reinforced with their gray-scale conversions as well as the original color images. Moreover, we generate and modify the training data by contrast-jitter (not color) and other conventional means of data augmentation. Beginning from the initial model trained with the massive amount of data, we utilize the NIR-chromaticity images with the ground-truth segmentation masks in our RANUS database to fine-tune the network. We find that this process mitigate the illumination variations between the visible and NIR spectrums, leading our NIR-chromaticity images to be used with the massive amount of RGB datasets.

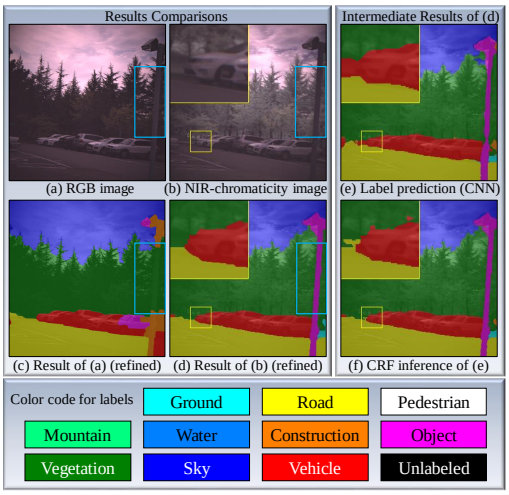


Fig. 4. Validation of our NIR-chromaticity image for scene parsing network DeepLab-CRF [4]. (a): RGB image. (b): NIR-chromaticity image. (c): Scene parsing result of (a). (d): Scene parsing result of (b). In (e) and (f), we show intermediate results of (d).

VI. PSEUDO COLORIZATION OF NIR IMAGE

After validating the effectiveness of the NIR-chromaticity images, we present a method which automatically generates the chromaticity from only the NIR image. To generate the NIR-chromaticity image, we use a gray-image colorization network as a baseline network and fine-tune the network with our own database. In this paper, we refer to this generated image as a *pseudo NIR-chromaticity* image.

Since the pixel-wise alignment of the RGB-NIR pair images in our database is accurate, we can directly use them for the deep learning architecture. Our architecture uses a recent colorization network [3]. Given that the original purpose of this method is to colorize a gray-scale (visible spectrum) image and not a NIR image, in our work we exploit their network and suitably train the model with our database. The network architecture uses eight repeated CNN layers without pooling. Compared to the baseline method which uses single RGB image, we use RGB and NIR pair images for the training. Given an input NIR image $\mathbf{X}_N \in \mathbb{R}^{h \times w \times 1}$ (where h is the image height and w is the image width), our goal is to train a mapping function between the NIR image $\mathbf{X}_N$ and the a, b chromaticity $\mathbf{Y}_{C,gt}$, that satisfies $\mathbf{Y}_{C,gt} = f(\mathbf{X}_N)$. To train the model, our $L2$ loss function is defined as follows:

$$L = \frac{1}{2} \sum_{i=1}^{h} \sum_{j=1}^{w} \left\| \mathbf{Y}_{C,gt}^{i,j} - \mathbf{Y}_{C}^{i,j} \right\|_2^2 , \tag{2}$$

where, $\mathbf{Y}_{C,gt}$ is the ground-truth chromaticity prepared from the RGB image and $\mathbf{Y}_C$ is the estimated a, b chromaticity after going through the network from the input NIR image. To represent the color information, we adopt the quantized ab output space in the baseline method. By quantizing the color gamut with 313 bins, the problem can be cast as a classification problem with the softmax classification loss in the network.

We observe that our NIR image roughly shares much information with the gray image. Therefore, instead of training the network from scratch, the initial model in [3] is a reasonable initialization for our problem. Beginning with this initial model which was trained with the ImageNet database, we fine-tune the network with our RGB-NIR database. For the

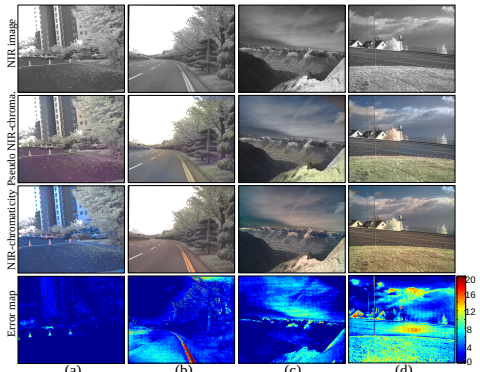


Fig. 5. Our NIR colorization results. We see our colorized NIR (2nd row) shows comparable result with the NIR-chroma. (3rd row)

training dataset, we use 38k RGB-NIR paired images from our database. For the testset, we use 1k image pairs evenly collected from each categories. We also validate our method on the EPFL benchmark database.

VII. Experimental Results

A. Experimental Setup

To measure the accuracy of our image-aligning method, we measured bad pixel rate and the root mean square error (RMSE) values using the Middlebury dataset. The bad pixel rate denotes the percentage of pixels that have fewer than ten color level errors (out of 255 levels, 8-bit). Compared to homography-based warping which yields 9 color level error and 14% bad pixels, our method is more accurate yielding only one color level error and 3% bad pixels.

For both colorization and scene parsing, we utilized Caffe [36] to train CNNs with two i7-4790 CPUs and NVIDIA GeForce GTX 1080, Titan X GPU. We conducted tests on the same PC with NVIDIA GeForce GTX 1080 GPU. To generate the NIR-chromaticity and to estimate the pseudo NIR-chromaticity, a time of less than 1 sec was required. Without image saving, these processes can be run in real-time. For scene parsing, the run time depends on the base networks. For example, the VGG-based network is less than 1 sec, whereas ResNet required more than 20 secs (including CRF post-processing).

B. NIR Colorization Results

Qualitative analysis: For the NIR image colorization step, we trained the network model with 235k iterations. In Figure 5, we demonstrate that our colorization for the NIR image works for various scenes. From the first row, the NIR image (single-channel), pseudo NIR-chromaticity image (colorized from the single-channel NIR image), ground- truth NIR-chromaticity image (derived using both cameras) are displayed. (a,b): Road scenes from our database. Green denoting vegetations, and dark gray representing roads are recovered in our pseudo NIR-chromaticity image. (c,d): A mountain view and a street dataset from [2]. In the 4th row, error maps of our pseudo NIR-chromaticity image against the ground truth are shown. The error map in a single color is computed by averaging the a and b channels in Lab space for every pixels.

Quantitative evaluation: In Table I, the performance of our NIR colorization method is evaluated using three different

TABLE I
QUANTITATIVE EVALUATION ON **COLORIZATION** (OURS VS [3]).

O/B	*Urban*	*Rural*	*Evening*	*Campus*	*EPFL*
ME	**6.4**/9.5	**4.5**/6.0	**2.1**/4.8	**2.8**/6.4	**2.6**/6.3
R5	**61.7**/61.5	**71.8**/71.7	**85.1**/68.2	**75.8**/54.2	**83.2**/71.6
R20	**88.6**/81.2	**93.1**/88.7	**97.6**/93.2	**98.3**/92.7	**96.7**/88.8

O/C	*Daimler*
ME	**7.4**/13.9
R5	**48.8**/42.1
R20	**88.8**/67.6

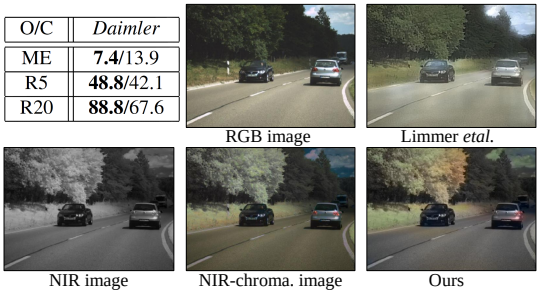


Fig. 6. Result comparison of ours and [20].

datasets. First, 1k frames of our database are used. To report the evaluation score, we categorize the test images into four categories reflecting their image characteristics. These are *Urban*, *Rural*, *Evening* and *Campus*. Secondly, we use a benchmark database from EPFL [1]. This database contains various outdoor sceneries. As error measures, we evaluate ME, R5 and R20. The ME represents the mean absolute error of the pixel intensity which ranges from 0 to 255 (the lower the better), and R5 and R20 are the percentages (%) of pixels having a color difference of less than 5 and 20 respectively (the higher the better). We found that the proposed fine-tuning method works well, yielding lower ME values than the baseline colorization model in [3]. The R5 and R20 measures also show higher scores on both our and the EPFL database. Additionally, we used Daimler [20] dataset shown in Fig.10 in the paper. Using this dataset, we compared our method with a concurrent method Limmer *et al.* [20]. In Figure 6, the results of the comparison are visualized and evaluated. The result from Limmer is compared with RGB image, and the result from our method is compared with NIR chroma. image. We see Limmer's shows blurry colorized result, which yields ME of 13.9. On the other hand, ours yields 7.4. Also the R5 and R20 measures of our method have 48.8 and 88.8% respectively, demonstrating that the results with the proposed method yielded higher accuracy as compared to that of the competitive method.

C. Scene Parsing Results

Evaluation on a benchmark: Motivated by the algorithm and dataset in [1], we extend the NIR imaging to a recent deep learning framework for vehicle applications. So we first evaluate our method on the EPFL dataset [1]. In Figure 7 (a), the RGB image parsing result does not separate OBJECT such as road signs hidden in VEGETATION, and does not locate MOUNTAIN. In contrast, our (pseudo) NIR-chromaticity images detect hidden objects and the hazy mountains. In Figure 7 (b,c), we find haze-enveloped mountains over water in the RGB image, whereas the mountains are clearer in the NIR image. Moreover, our pseudo NIR-chromaticity image better segments the SKY, WATER, and MOUNTAIN.

The quantitative evaluation result with this database is shown in Table II. Compared to our database, which targets

TABLE II
EVALUATION USING **BENCHMARK DATASET [1]** WITH IOU METRIC (JACCARD INDEX). OVERLAPPED CLASSES ARE EVALUATED.

Class	RGB	NIR-chroma.	NIR	Pseudo
sky	87.57	87.80	83.66	87.52
ground	32.58	35.65	22.58	33.94
water	52.27	53.96	50.71	52.39
road	77.67	77.79	68.17	76.83
construction	65.69	67.24	59.62	67.11
vegetation	79.39	82.19	73.21	81.47
object	28.84	29.57	24.81	27.87
mean IOU	60.57	**62.03**	54.68	**61.02**

TABLE III
QUANTITATIVE COMPARISON OF **OURS** (DEEP) AND [1] (HAND-CRAFT).

RGB Camera Only			NIR Camera Only		
Col_rgb	Sift_rgb	Deep_rgb	Sift_nir	Deep_nir	Deep_nirc
59.21	60.33	**60.57**	51.12	**54.68**	**61.02**

dynamic road scenes, this dataset focuses on static scenes. Therefore, the class definitions of the two databases differ slightly. For a fair comparison, we re-labeled the training data in accordance with the benchmark dataset and re-trained the deep scene parsing network. In Table II, we find that our NIR-chroma. image yields a higher IOU (62.03) than the RGB image (60.57) in a comparison of two 3-channel images. We also find that our colorization of an NIR image boosts the mean IOU value from 54.68 to 61.02, when comparing our pseudo NIR-chroma. image with a single-channel NIR image. Note that the EPFL dataset does not have an adequate number of ground truth masks for finetuning.

Comparison with the conventional approach: Our method is compared with a conventional hand-craft feature-based scene parsing method [1]. This work evaluate the accuracy of scene parsing with different hand-craft features such as SIFT. In Table III, the IOU comparison result is shown. First, we compare the result when using an RGB image. Compared to *Col_rgb* and *Sift_rgb*, the *Deep_rgb* yields higher IOU values. This demonstrates the advantage of using deep features for the scene parsing as compared to using the conventional SIFT features. Second, we compare the results in the NIR domain. When using the SIFT feature of the NIR images, the IOU score is 51.12. On the other hand, when the deep features are extracted and used in the NIR image, the IOU score is improved to 54.68. Finally, we find that the accuracy is enhanced even more when the single-channel NIR image automatically generates the chromaticity and this pseudo NIR-chromaticity is used. The IOU of this case (*Deep_nirc*) is 61.02, which is a significant improvement over the other methods. This demonstrates that our colorization of NIR images is effective.

Evaluation on our database: Figure 7 (d-g) show examples of the comparison results on our database. The parsing results are especially highlighted in challenging scenes, such as those with low-light levels, ambiguous contrasts, and poor visibility conditions in the RGB images: (d) In the RGB image, the fences, building and vegetations are segmented incorrectly due to poor contrast. In contrast, CONSTRUCTION and VEGETATION in our (pseudo) NIR-chromaticity images are clearly distinguished, and the fences are accurately detected. (e) In the park, a dark vehicle is approaching and a person wearing dark clothing is walking across it, a potentially dangerous situation. In the RGB image, the colors of the car and the pedestrian are similar, and parsing fails. However, the NIR image captures the clear contrast between PEDESTRIAN and VEHICLE with VEGETATION. The parsing results are enhanced with the pseudo NIR-chromaticity image for SKY and GROUND owing to our colorization approach. (f) In the RGB image, a preceding black car is hidden in the bushes. Thus, the rear-end of VEHICLE is not segmented accurately, and the bush area of VEGETATION is incorrectly identified as a road region. We find that this problem is solved with the NIR information, and the chromaticity information is helpful for better segmentation of the ROAD. (g) In poor weather with haze, a car is driving along a rural road. In the RGB result, the mountain and road sign are incorrectly segmented. However, our (pseudo) NIR-chromaticity image successfully distinguishes the MOUNTAIN, VEGETATION, OBJECT, and CONSTRUCTION classes in the image.

TABLE IV
QUANTITATIVE EVALUATION USING **OUR DATASET**.

Class	RGB	NIR-chroma.	NIR	Pseudo
sky	93.62	93.63	91.22	93.08
ground	35.26	50.44	25.64	47.41
water	24.00	31.24	15.01	46.51
mountain	26.48	42.42	34.44	32.61
road	87.61	76.96	88.71	87.51
construction	43.26	64.10	46.73	53.61
vegetation	66.54	82.78	74.26	78.76
object	21.08	29.98	19.56	21.84
vehicle	53.50	66.75	50.09	64.95
person	22.36	30.81	25.13	26.96
mean IOU	47.37	**56.91**	47.08	**55.32**

In Table IV, we measure the IOU for each of the defined ten classes. Among the various modalities, we found that our NIR-chromaticity image yields a higher mean IOU (56.91) as compared to a normal RGB image (47.37) in most classes. Moreover, our pseudo colorized NIR image yields a higher IOU (55.32) than the single-channel NIR approach (47.08) comparable to the outcome with the NIR-chromaticity image. This demonstrates that the inferred chromaticity is beneficial when used for deep scene parsing. We also validated that our colorization of NIR images can boost the accuracy of scene parsing even when using the only a single-channel NIR image. Specifically, it shows that the IOU value is much higher in the classes of CONSTRUCTION (53.61), VEGETATION (78.76), VEHICLE (64.95), PERSON (26.96), where challenging scenes often arise.

Table V summarizes the results of the comparison. In order to analyze the key factors in the presented algorithm, we tested all the CNN models with different conditions and image modalities. These steps include gray-intensity jitter (JIT), iterative post processing (PP), and fine-tuning (FT) using our database. For every image modality, we use the two base networks: the VGG-network and the residual (Res)-network. We found that the residual network outperforms in all test cases. From the baseline models initialized with the same images, we perform additional processings and evaluate the IOU of each step. Our technique of using pseudo colors outperformed the method which used only NIR images for all of the trained models. Moreover without fine-tuning, the positive

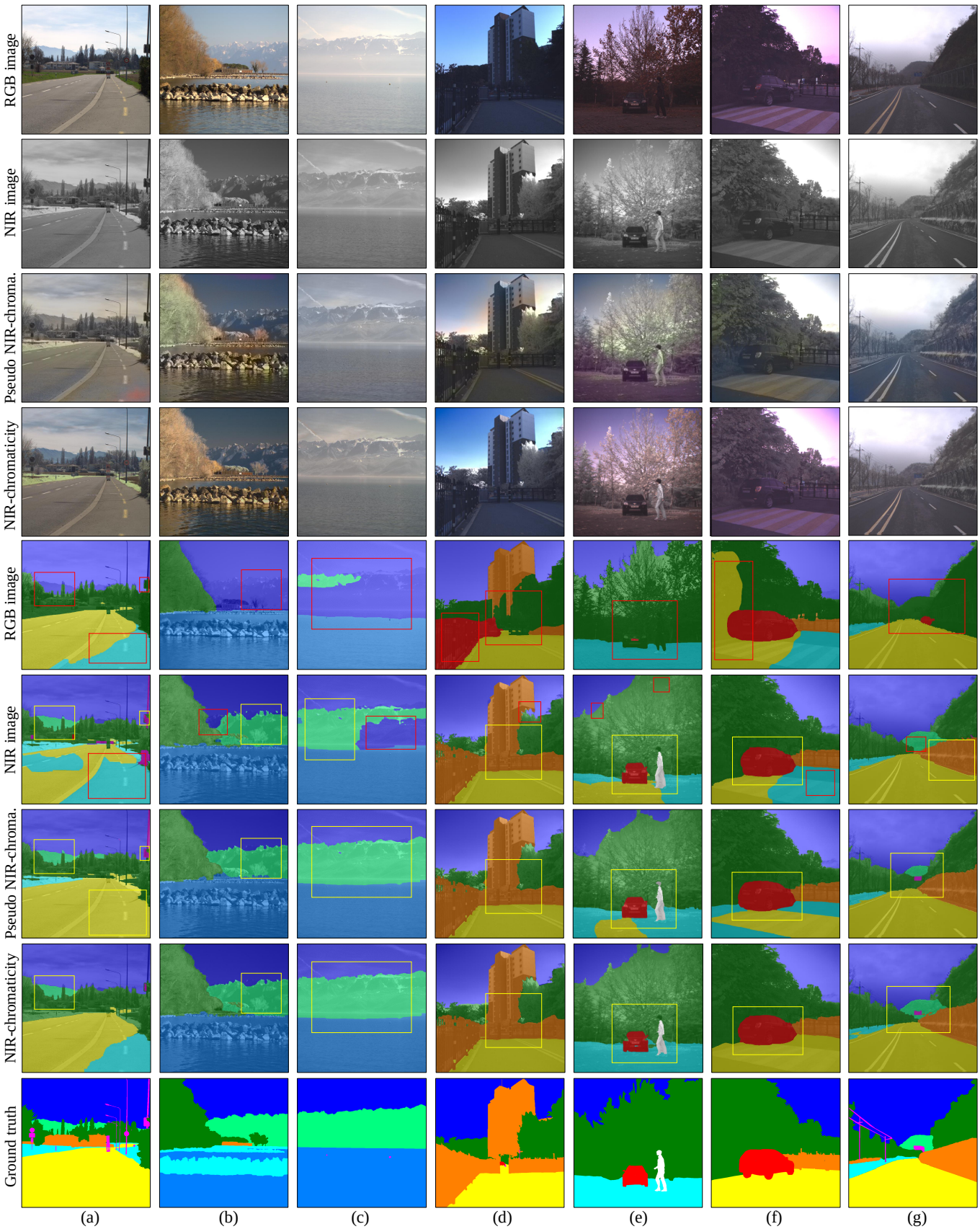


Fig. 7. Scene parsing results. (1-4) rows: RGB image, NIR image, pseudo NIR-chromaticity image and NIR-chromaticity image respectively, (5-8) rows: Scene parsing results of (1-4) respectively. Note that the parsing ambiguities in conventional RGB images, shown as red boxes, are accurately resolved in our result (yellow boxes). See the color code for class labels in Figure 4

TABLE V
COMPARISON RESULT OF DIFFERENTLY TRAINED NETWORK MODELS.

	Trained model	RGB	NIR-chroma.	NIR	Pseudo
VGG-Net	RGB	39.12	38.05	36.93	37.20
	RGB+JIT	38.77	38.77	37.35	37.64
	RGB+JIT+PP	38.96	39.02	37.33	37.71
	RGB+JIT+FT	37.34	41.77	36.63	39.57
Res-Net	RGB	47.37	46.61	44.92	45.25
	RGB+JIT	45.93	48.66	47.08	47.79
	RGB+JIT+PP	46.52	49.85	46.71	48.18
	RGB+JIT+FT	46.94	**56.91**	39.30	**55.32**

impact of utilizing the estimated chromaticity channels was highlighted in the RGB+JIT+PP case. Regardless of the base models (VGG, ResNet), we observe that the inferred chromaticity channels are informative. Lastly, we further improve our results using our own database, RANUS. The pseudo images (55.32) are working closely to NIR-chromaticity images (56.91) nonetheless they are actually derived from the NIR-chromaticity trained model. For the fine-tuning step, the training (3k) and test (1k) images are evenly selected from each category in our database. We separately select 10% of the training images for validation during the training process. There is no correlation between the training, validation, and test images and no bias for a particular category.

VIII. CONCLUSION

In this work, we have built a large RGB-NIR video pair database with pixel-level segmentation masks. The pairwise frames were precisely aligned using a multi-spectral alignment approach. By using our database into the scene parsing network, we validated that augmenting the RGB with the NIR brightness is beneficial for scene parsing. We also trained a spectrum relation between the NIR image intensity and the color component from the RGB image by deeply training a colorization network. By doing so, we were able to generate a pseudo NIR-chromaticity image and demonstrated it helps scene parsing under challenging conditions where a normal RGB image often fails. For semantic segmentation of non-RGB images, we expect our database to be expanded toward including more time and place variations.

REFERENCES

[1] N. Salamati, D. Larlus, G. Csurka, and S. Süsstrunk, "Semantic image segmentation using visible and near-infrared channels," in *ECCV Workshops and Demonstrations*. Springer, 2012, pp. 461–471.

[2] M. Brown and S. Süsstrunk, "Multi-spectral sift for scene category recognition," in *CVPR*, 2011, pp. 177–184.

[3] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," *arXiv preprint arXiv:1603.08511*, 2016.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv preprint arXiv:1606.00915*, 2016.

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[6] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[7] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *CVPR*, 2014, pp. 891–898.

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE CVPR*, 2016.

[9] N. Salamati, Z. Sadeghipoor Kermani, and S. Susstrunk, "Analyzing Near-infrared Images for Utility Assessment," in *IS&T/SPIE Electronic Imaging: Human Vision and Electronic Imaging*, 2011.

[10] H. Tang, X. Zhang, S. Zhuo, F. Chen, K. N. Kutulakos, and L. Shen, "High resolution photography with an rgb-infrared camera," in *IEEE ICCP*, 2015, pp. 1–10.

[11] D. Rufenacht, C. Fredembach, and S. Susstrunk, "Automatic and accurate shadow detection using near-infrared information," *IEEE TPAMI*, vol. 36, no. 8, pp. 1672–1678, 2014.

[12] G. Choe, J. Park, Y.-W. Tai, and I. S. Kweon, "Refining geometry from depth sensors using ir shading images," *IJCV*, vol. 122, no. 1, pp. 1–16, 2017.

[13] G. Choe, S. G. Narasimhan, and I. S. Kweon, "Simultaneous estimation of near ir brdf and fine-scale surface geometry," in *IEEE CVPR*, 2016, pp. 2452–2460.

[14] C. Feng, S. Zhuo, X. Zhang, L. Shen, and S. Süsstrunk, "Near-infrared guided color image dehazing," in *IEEE ICIP*, 2013, pp. 2363–2367.

[15] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," in *ACM TOG*, vol. 23, no. 3, 2004, pp. 689–694.

[16] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," *arXiv preprint*, 2016.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint*, 2014.

[18] C. Fredembach and S. Süsstrunk, "Colouring the near-infrared," in *Color and Imaging Conference*, vol. 2008, no. 1. Society for Imaging Science and Technology, 2008, pp. 176–182.

[19] C.-H. Son, X.-P. Zhang, and K. Lee, "Near-infrared coloring via a contrast-preserving mapping model," in *Signal and Information Processing, 2015 IEEE Global Conference on*, 2015, pp. 677–681.

[20] M. Limmer and H. Lensch, "Infrared colorization using deep convolutional neural networks," *arXiv preprint arXiv:1604.02245*, 2016.

[21] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *IEEE CVPR*, 2014.

[22] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *ECCV*. Springer, 2014, pp. 297–312.

[23] B. Fulkerson, A. Vedaldi, S. Soatto, *et al.*, "Class segmentation and object localization with superpixel neighborhoods." in *IEEE ICCV*, vol. 9. Citeseer, 2009, pp. 670–677.

[24] V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Adv. Neural Inf. Process. Syst*, 2011.

[25] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE TPAMI*, vol. 35, no. 8, 2013.

[26] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[27] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *IEEE ICCV*, 2015, pp. 2650–2658.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[29] N. Salamati and S. Süsstrunk, "Material-based object segmentation using near-infrared information," in *Color and Imaging Conference*. Society for Imaging Science and Technology, 2010.

[30] S. M. Directorate, "Reflected near-infrared waves," *Mission:Science*, 2010.

[31] A. T. Young, "Rayleigh scattering," *Applied Optics*, vol. 20, no. 4, pp. 533–535, 1981.

[32] S. Kim, D. Min, B. Ham, S. Ryu, M. N. Do, and K. Sohn, "Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence," in *IEEE CVPR*, 2015, pp. 2103–2112.

[33] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE TPAMI*, vol. 33, no. 5, pp. 978–994, 2011.

[34] Q. Yang, "A non-local cost aggregation method for stereo matching," in *IEEE CVPR*. IEEE, 2012, pp. 1402–1409.

[35] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM TOG*, vol. 23, no. 3, 2004, pp. 309–314.

[36] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.