Effective Source-Aware Domain Enhancement and Adaptation for CNN-Based Object Segmentation

KUO-LIANG CHUNG[1], (Senior Member, IEEE), YA-YUN CHENG[1]
[1]Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei 10672, Taiwan.
Corresponding author: Kuo-Liang Chung (e-mail: klchung01@gmail.com).
This work was supported by Grants MOST-107-2221-E-011-108-MY3 and MOST-108-2221-E-011-077-MY3.

ABSTRACT In this paper, we propose an effective source-aware domain enhancement and adaptation (SDEA) approach to increase the accuracy of the existing convolutional neural network-based (CNN-based) object segmentation methods. We first scoop out the source elements, such as the falling-leaves, manhole covers, cirrus clouds, and advertisements, which often cause invalid object segmentation and make the existing object segmentation methods provide unreliable information to the ADAS (automatic driving assistance systems) applications. Secondly, we create a new GTA5-like (Grand Theft Auto V-like) dataset with the scenarios including these source elements. Furthermore, we perform a domain adaptation on the created GTA5-like dataset to generate a photo-realistic GTA5-like dataset, namely GTA5_s^{SDEA}. Without the need to relabel the pixel-annotations for GTA5_s^{SDEA}, we combine GTA5_s^{SDEA} with the realistic dataset, namely Camvid, to constitute a newly enhanced dataset. After retraining the existing CNN-based object segmentation methods by using our enhanced dataset, it can achieve substantial segmentation accuracy improvement. The comprehensive experimental results have demonstrated the clear accuracy improvement merit by applying our SDEA approach to the state-of-the-art object segmentation methods on FCN (Fully Convolutional Networks), SegNet-basic, AdaptSegNet, and Gated-AdaptSegNet, providing more reliable information to ADAS applications.

INDEX TERMS ADAS (automatic driving assistance systems), CNN (Convolutional Neural Networks), Domain Adaptation, Domain Enhancement, GTA5 (Grand Theft Auto V), mIoU (mean intersection over union), Object Segmentation Accuracy.

I. INTRODUCTION

Recently, developing object segmentation methods using convolutional neural networks (CNN) [1], [4], [16]–[18], [21], [27], [32] has received great attention in different applications, particularly in ADAS (automatic driving assistance systems) applications [1], [24]. In ADAS applications, for each image frame, the CNN-based object segmentation method often considers up to nineteen object types, namely road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle, and bike. Due to the high object segmentation accuracy demand in ADAS applications, designing a novel approach to improve the existing CNN-based object segmentation methods is an important task.

In the past years, by using the CNNs, animation-based dataset enhancement, and domain adaptation, some successful object segmentation methods have been developed to increase the segmentation accuracy, providing more reliable information to ADAS applications in lane detection [7], [14], traffic sign recognition [16], departure/collision warning [26], [29], and vanishing point detection [3], [13]. In the next subsection, the related work is introduced.

A. RELATED WORK

Based on the end-to-end fully-convolutional network (FCN) model, in which there are 15 convolutional layers in the encoder and there are three deconvolutional layers in the decoder, an effective FCN-based object segmentation method [18], [25] was proposed. The configuration of FCN used in their method is shown in Table 1. To achieve higher object segmentation accuracy, Badrinarayanan *et al.* [1] modified the FCN model by reducing the number of convolutional layers from 15 layers to nine, and replacing the three deconvolutional layers by five upsampling layers. Their modified FCN model is called SegNet-basic. Table 2 shows the configuration of SegNet-basic. Due to the available source

TABLE 1. The configuration of FCN.

Layer	Filter (#Filters)	Feature Map
Conv1	3x3x3 (64)	256x512x64
Conv2	3x3x64 (64)	256x512x64
Maxpool	-	128x256x64
Conv3	3x3x64 (128)	128x256x128
Conv4	3x3x128 (128)	128x256x128
Maxpool	-	64x128x128
Conv5	3x3x128 (256)	64x128x256
Conv6	3x3x256 (256)	64x128x256
Conv7	3x3x256 (256)	64x128x256
Maxpool	-	32x64x256
Conv8	3x3x256 (512)	32x64x512
Conv9	3x3x512 (512)	32x64x512
Conv10	3x3x512 (512)	32x64x512
Maxpool	-	16x32x512
Conv11	3x3x512 (512)	16x32x512
Conv12	3x3x512 (512)	16x32x512
Conv13	3x3x512 (512)	16x32x512
Maxpool	-	8x16x512
Conv14	3x3x512 (4096)	8x16x4096
Conv15	3x3x4096 (19)	8x16x19
DeConv1	4x4x19 (19)	16x32x19
Fuse	-	16x32x19
DeConv2	4x4x19 (19)	32x64x19
Fuse	-	32x64x19
DeConv3	16x16x19 (19)	256x512x19
Softmax	-	256x512x19

codes, the object segmentation methods by using FCN and SegNet-basic are included in the comparative methods to justify the segmentation accuracy improvement merit of our source-aware domain enhancement and adaptation (SDEA) approach.

To increase the segmentation accuracy, researchers have put much effort into capturing realistic images, and then they labeled each pixel annotation for these captured images for creating realistic datasets, such as Camvid [2], Cityscapes [5], KITTI [9], and Urban LabelMe [23]. However, capturing and labeling more realistic images to enhance the datasets is quite expensive and time-consuming. To solve this expensive and time-consuming problem, Richter *et al.* [20] proposed a cheap and effective animation-based approach to create a synthetic GTA5 dataset with 7500 synthetic images. Experimental data demonstrated that using this synthetic GTA5 dataset and one realistic dataset, e.g. Camvid, as the hybrid training set, the CNN-based object segmentation accuracy can be improved.

Although Richter *et al.*'s animation-based dataset enhancement approach [20] can improve the segmentation accuracy relative to the traditional approach, due to the domain shift problem between the realistic dataset, namely Camvid, and the synthetic dataset, namely GTA5, the segmentation accuracy using the hybrid training dataset "GTA5+Camvid" to train the CNN-based frameworks is not as good as expected.

TABLE 2. The configuration of SegNet-basic.

Layer	Filter (#Filters)	Feature Map
Conv1	7x7x3 (64)	256x512x64
Maxpool	-	128x256x64
Conv2	7x7x64 (64)	128x256x64
Maxpool	-	64x128x64
Conv3	7x7x64 (64)	64x128x64
Maxpool	-	32x64x64
Conv4	7x7x64 (64)	32x64x64
Maxpool	-	16x32x64
Upsample	-	32x64x64
Conv5	7x7x64 (64)	32x64x64
Upsample	-	64x128x64
Conv6	7x7x64 (64)	64x128x64
Upsample	-	128x256x64
Conv7	7x7x64 (64)	128x256x64
Upsample	-	256x512x64
Conv8	7x7x64 (64)	256x512x64
Conv9	1x1x64 (19)	256x512x19
Softmax	-	256x512x19

To solve this domain shift problem, several domain adaptation approaches were proposed to reduce the gap between the synthetic dataset and the realistic dataset. Tsai *et al.* [27] proposed a generated adversarial network-based (GAN-based) object segmentation method, called the AdaptSegNet method, that used the adversarial training to align pixel-level ground truth in the output space. Based on GAN, Lin *et al.* [17] proposed the Gated-AdaptSegNet based method that used a foreground adaptation module to separate the foreground and background for improving the segmentation accuracy.

Different from Tsai *et al.*'s approach [27], Zhang *et al.* [33] transformed the GTA5 dataset to a photo-realistic dataset, denoted by $GTA5_s$, by using the style transfer technique. Experimental data illustrated that using the hybrid training dataset "$GTA5_s$+Camvid" as the enhanced training dataset can increase the object segmentation accuracy. Due to the available codes, the AdaptSegNet and Gated-AdaptSegNet methods are included in the comparative methods to justify the accuracy improvement by using our SDEA approach.

B. MOTIVATION

For convenience, let the FCN-based object segmentation method, the SegNet-basic-based object segmentation method, the AdaptSegNet-based object segmentation method, and the GatedAdaptSegNet object segmentation method be denoted by FCN, SegNet-basic, AdaptSegNet, and GatedAdaptSegNet, respectively. Based on the enhanced dataset "$GTA5_s$+Camvid", after training the above-mentioned four object segmentation methods, we found that in the testing step, the two sources, namely falling-leaves

and manhole covers, often cause invalid road segmentation; another two sources, namely cirrus clouds and advertisements, often cause invalid sky and building segmentation, respectively. In particular, the invalidly segmented road, sky, and building information may result in improper decisions in ADAS applications. For example, the invalid road segmentation may lead to improper lane detection, invalid traffic sign recognition, and wrong departure/collision warning; the invalidly segmented sky may lead to incorrect vanishing point detection.

Without the need to relabel the pixel-annotations, the above source-aware observation prompted us to develop a novel and effective source-aware domain enhancement and adaptation (SDEA) approach to create a newly enhanced dataset "GTA5$_s^{SDEA}$". And then, based on our dataset GTA5$_s^{SDEA}$, the retrained version of the four considered object segmentation methods, namely FCN, SegNet-basic, AdaptSegNet, and Gated-AdaptSegNet, can have higher segmentation accuracy. Note that our SDEA approach is infeasible to retrain the Mask R-CNN based object segmentation method [12] because among the eighty objects considered by Mask R-CNN, only nine, namely person, bicycle, car, motorcycle, bus, train, truck, traffic light, and stop sign, are useful in ADAS applications. Therefore, we do not apply our SDEA approach to the Mask R-CNN model.

C. CONTRIBUTION

To overcome the above-mentioned weakness and limitation existing in the related work, this paper proposes a novel and effective SDEA approach to achieve substantial object segmentation accuracy improvement for the four considered CNN-based object segmentation methods. The three contributions of our SDEA approach are clarified as follows.

In the first contribution, the proposed SDEA approach scoops out the sources, namely the falling-leaves, manhole covers, cirrus clouds, and advertisements, which infrequently or irregularly appear in the real situation but often cause invalid object segmentation; the invalid object segmentation information tends to interfere with incorrect decisions in ADAS applications. Therefore, we propose a source-pasting technique to create a new GTA5-like dataset which contains the scenarios including these sources. In each GTA5-like image, the additive sources come from the sub-image cutting off from a realistic image in the dataset "Camvid".

In the second contribution, we perform a domain adaptation on our GTA5-like dataset to generate a photo-realistic GTA5-like dataset, called "GTA5$_s^{SDEA}$". Accordingly, the new hybrid dataset, called "GTA5$_s^{SDEA}$+Camvid," is created. Due to inheriting the originally labeled pixel annotation in GTA5$_s$ and Camvid, the labeling work on GTA5$_s^{SDEA}$ can be waived, exempting the labeling-time overhead. Furthermore, we apply our new hybrid dataset to retrain the four considered object segmentation methods, namely FCN, SegNet-basic, AdaptSegNet, and Gated-AdaptSegNet, achieving substantial object segmentation accuracy improvement.

In the third contribution, the comprehensive experimental data have confirmed that our SDEA approach with our newly enhanced dataset "GTA5$_s^{SDEA}$+Camvid" can substantially improve the object segmentation accuracy for the above-mentioned four considered CNN-based object segmentation methods. In terms of mean intersection over union (mIoU) to measure the object segmentation accuracy for the considered nineteen objects, the mIoU gains of our SDEA approach over FCN, SegNet-basic, AdaptSegNet, and Gated-AdaptSegNet are 1.1, 3.1, 1.5, and 1.7, respectively, providing more reliable object segmentation information to ADAS applications, making more trustworthy traffic decisions.

The rest of this paper is organized as follows. Section II presents our SDEA approach and describes how to build up the newly enhanced dataset "GTA5$_s^{SDEA}$+Camvid". Section III reports the object segmentation accuracy improvement merit of our SDEA approach relative to the four state-of-the-art CNN-based object segmentation methods. Section IV addresses some concluding remarks.

II. THE PROPOSED SDEA APPROACH

We first scoop out sources causing invalid object segmentation in the testing step. Then, without the pixel-annotation labeling overhead, a source-pasting technique is proposed to create an enhanced version of the dataset "GTA5$_s$," called "GTA5$_s^{SDEA}$," which contains the scenarios including these sources. Furthermore, we create a newly enhanced dataset "GTA5$_s^{SDEA}$+Camvid" which will be used to retrain the above-mentioned CNN-based object segmentation methods for increasing their segmentation accuracy.

A. SCOOP OUT SOURCES CAUSING INVALID OBJECT SEGMENTATION

From the observation of the object segmentation results in the testing step, we found that some invalid segmentation for objects, such as road, sky, and buildings, is often caused by the sources, namely the falling-leaves, manhole covers, cirrus clouds, and advertisements, because these sources infrequently or irregularly appear in the testing images. In particular, these invalid segmented roads, sky, and buildings may provide wrong information to ADAS applications in lane detection, departure/collision warning, and vanishing point detection.

Before taking practical examples to explain why the above-mentioned sources cause invalid object segmentation, the loss function $Loss(I)$ used for object segmentation is defined by

$$Loss(I) = -\sum_{c \in C} Y^{c(I)} log(S^{c(I)}) \tag{1}$$

where $I (\in \mathbb{R}^{H \times W \times 3})$ denotes the input H $\times$ W RGB full-color image; $S^{c(I)} \in \mathbb{R}^{H \times W \times C}$ denotes the output H $\times$ W binary map, in which C denotes the set of all object classes and $S^{c(I)}$ denotes the resultant segmentation map for the object class $c \in C$. When the entry in $S^{c(I)}$ is 1, it indicates that the recognized object class for that pixel is equal to the

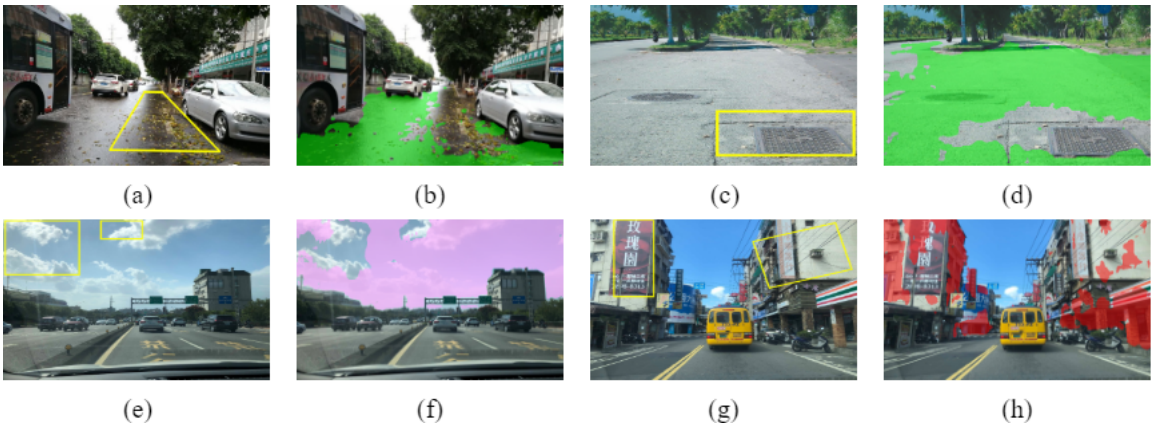


FIGURE 1. Four sources causing invalid object segmentation. (a) Falling-leaves. (b) Invalid road segmentation due to (a). (c) Manhole cover. (d) Invalid road segmentation due to (c). (e) Cirrus clouds. (f) Invalid sky segmentation due to (e). (g) Advertisements. (h) Invalid building segmentation due to (g).

object class c; otherwise, it denotes the wrong recognition for the pixel. $Y^{c(I)}$ denotes the ground-truth labeled annotation map.

Based on the dataset "GTA5$_s$+Camvid" on FCN, in which the photo-realistic dataset GTA5$_s$ is obtained by performing the domain adaptation method "Photorealistic Image Stylization [15]" on the synthetic dataset GTA5, we take four practical testing images to explain why the above-mentioned four sources lead to the invalid object segmentation problem, prompting us to propose the SDEA approach to solve this important problem.

As shown in Fig. 1(a), the falling-leaves surrounded by a yellow trapezoid on the lane cause an invalid road segmentation, as shown in Fig. 1(b). As shown in Fig. 1(c), the manhole cover on the lane surrounded by a yellow rectangle causes an invalid road segmentation, as shown in Fig. 1(d), because the texture of the manhole cover is different from that of the road. As for the cirrus clouds and advertisements shown in Fig. 1(e) and Fig. 1(g), respectively, the invalid segmented sky and buildings are illustrated in Fig. 1(f) and Fig. 1(h).

B. THE PROPOSED SOURCE-PASTING TECHNIQUE TO CREATE A NEWLY ENHANCED DATASET

In Fig. 1, four invalid object segmentation examples caused by four sources have been demonstrated. Capturing more realistic images with the scenarios containing the considered four sources is a straightforward way to enhance the dataset, but it is expensive and time-consuming. In addition, it is also quite time-consuming to label each pixel annotation of these captured real images. In what follows, without pixel-annotation labeling overhead, we propose a fast and effective source-pasting technique to create a new photo-realistic dataset, in which the scenarios contain these sources coming from the images in "Camvid," and then we combine it with "Camvid" to create the newly enhanced dataset.

For easy exposition of the proposed SDEA approach, we first explain how to create a new synthetic GTA5-like image containing the falling-leaves. Given a labeled synthetic GTA5 image in Fig. 2(a), from the dataset "Camvid," we select one real image containing falling-leaves, as shown in Fig. 2(b). Then, we paste the subimage containing falling-leaves, which is cut off from Fig. 2(b), to Fig. 2(a), creating the synthetic GTA5-like image shown in Fig. 2(c). In Fig. 2(c), all the pixels in the falling-leaves inherit the original labeled annotation in Fig. 2(b), and except for the falling-leaves, all the pixels in Fig. 2(c) inherit the original labeled annotation in Fig. 2(a), waiving the pixel-based labeling overhead.

After performing the style transform on the synthetic GTA5-like image via the domain adaptation method [15], the resultant photo-realistic GTA5-like image is shown in Fig. 2(d). As for the manhole cover case, by the same argument, Figs. 2(e)-(h) illustrate the corresponding four snapshots. Figs. 2(i)-(l) and Figs. 2(m)-(p) show the corresponding snapshots for the cirrus cloud and advertisement sources with respect to sky and building, respectively. However, using the above domain adaptation way to create the new photo-realistic GTA5-like images, the versatility of the created images is still not enough. To overcome this disadvantage and automatically generate more new photo-realistic GTA5-like images with different styles, we deploy the weather, namely the sunny day, the rainy day, and the overcast day, influence and the time period, the daytime, the nighttime, and the twilight, influence into the domain adaptation of our SDEA approach to increase the diversity of the newly created photo-realistic GTA5-like images as quick as possible.

By using our proposed SDEA approach, let the newly created photo-realistic GTA5-like dataset be denoted by GTA5$_s^{SDEA}$. Consequently, the newly enhanced dataset "GTA5^{SDEA}+Camvid" is used to retrain the four considered object segmentation methods, namely FCN, SegNet-basic, AdaptSegNet, and Gated-AdaptSegNet, to increase the segmentation accuracy, providing more reliable segmentation information to ADAS applications.

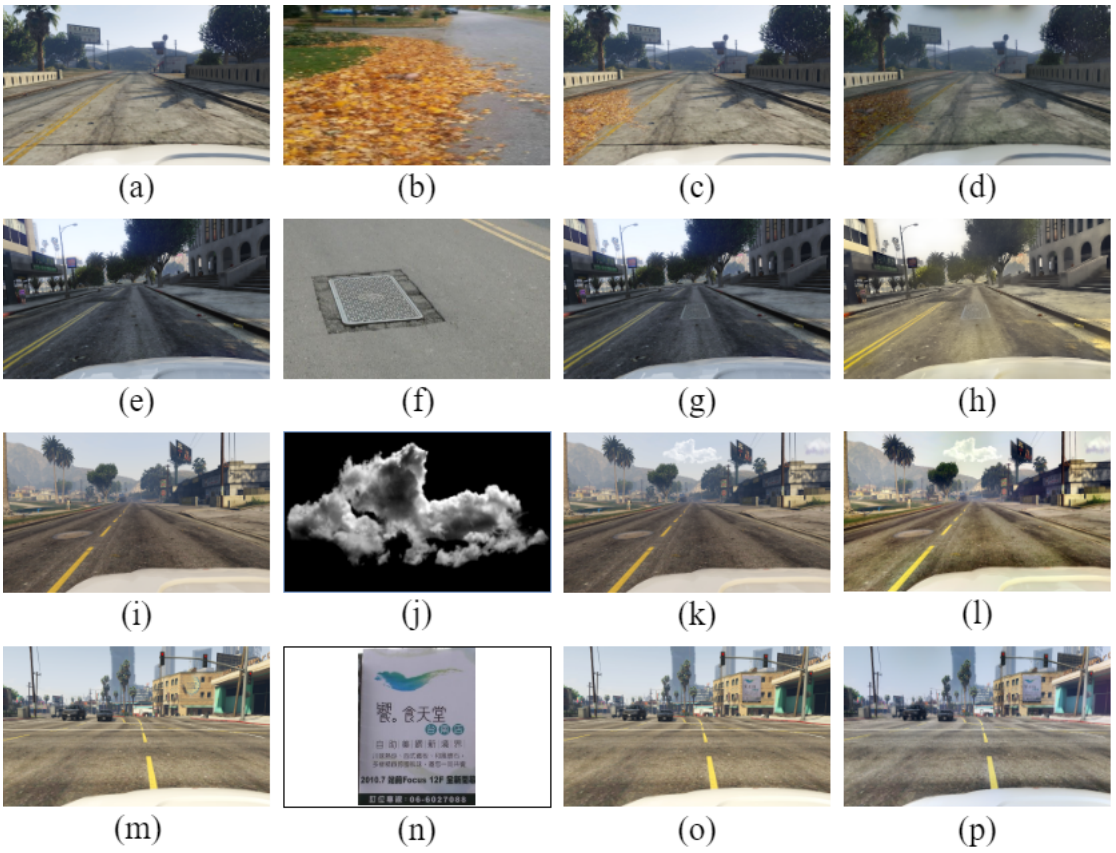


FIGURE 2. Four generated photo-realistic GTA5-like images by using our SDEA approach. (a) The first GTA5 image example. (b) A realistic image containing the falling-leaves source. (c) The synthetic GTA5-like image of (a). (d) The created photo-realistic GTA5-like image of (c). (e) The second GTA5 image example. (f) A realistic image containing the manhole-cover source. (g) The synthetic GTA5-like image of (e). (h) The created photo-realistic GTA5-like image of (g). (i) The third GTA5 image example. (j) A realistic image containing the cirrus cloud source. (k) The synthetic GTA5-like image of (i). (l) The created photo-realistic GTA5-like image of (k). (m) The fourth GTA5 image example (n) A realistic image containing the advertisements source. (o) The synthetic GTA5-like image of (m). (p) The created photo-realistic GTA5-like image of (o).

III. EXPERIMENTAL RESULTS

Based on the training dataset "Camvid+GTA5$_s$" with 1402 (= 701+701) images, as the four comparison baselines, four sets of experiments are carried out to show the object segmentation accuracy of the four comparative object segmentation methods , namely FCN [18], SegNet-basic [1], AdaptSegNet-based [27], and Gated-AdaptSegNet-based [17].

In order to demonstrate the accuracy improvement merit of our SDEA approach, we apply the proposed new dataset "Camvid+GTA5$_s^{SDEA}$" with 1722 (= 701+1021) images to train the above-mentioned four CNN-based object segmentation models. For convenience, the four retrained versions of the four object segmentation methods are called FCN^{SDEA}, SegNet-basic^{SDEA}, AdaptSegNet^{SDEA}, and Gated-AdaptSegNet^{SDEA} , respectively. For fairness, to compare the segmentation accuracy performance of all the considered object segmentation methods, we utilize the same testing dataset which consists of 580 images of which 500 are randomly collected from the dataset "Cityscapes" and 80 are captured from the real urban world and can be accessed from the website [8]. Note that in our two-step SDEA approach, the experimental results indicated that based on the dataset "Camvid+GTA5^{SDEA} ," the accuracy improvement effect of the first step, namely the source-aware based domain enhancement step, is incremental relative to the baseline models, while based on the dataset "Camvid+GTA5$_s^{SDEA}$," the accuracy improvement effect of our two-step SDEA approach, namely the source-aware based domain enhancement and adaptation, is obvious relative to the baseline models.

All experiments are implemented using a desktop with an Intel Core i7-7700 CPU running at 3.6 GHz with 32 GB RAM and an NVIDIA 1080Ti GPU. The operating system is Microsoft Windows 10 64-bit. The program development environment is PyCharm Professional with the Python programming language.

A. OBJECT SEGMENTATION ACCURACY IMPROVEMENT MERIT OF FCN^{SDEA}

In the first set of experiments, the "mIoU" gain is used to show the average object segmentation accuracy improvement merit of the proposed FCN^{SDEA} method over the FCN method [18].

The metric "IoU" is used to measure the object segmenta-

TABLE 3. The mIoU improvement merit of the proposed SDEA approach relative to the FCN method [18].

	road	sidewalk	building	wall	fence	pole	light	sign	veg	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
FCN	69.1	18.3	57.7	1.4	0.4	3.1	2.9	2.6	56.3	11.9	61.7	5.3	2.2	48.4	0.5	0.7	0.0	0.1	1.1	18.1
FCN^{SDEA}	71.1	24.1	58.8	3.0	1.0	3.8	3.0	3.4	56.1	11.2	64.4	5.2	2.7	53.6	0.7	0.9	0.0	0.0	1.0	19.2

tion accuracy of one object, and "IoU" is defined by

$$\text{IoU(object)} = \frac{|\text{Detected object pixels} \bigcap \text{Ground truth object pixels}|}{|\text{Detected object pixels} \bigcup \text{Ground truth object pixels}|} \quad (2)$$

In Eq. (2), "$\bigcap$" and "$\bigcup$" denote the "intersection" and "union" operations, respectively. The metric "mIoU" is used to measure the expected value of the "IoU" values for all considered objects.

In terms of "IoU" Table 3 tabulates the segmentation accuracy of each object among the considered 19 objects; the IoU value of each object is listed below the object field. The mIoU value of all the objects is listed in the final column of Table 3. Table 3 indicates that the mIoU gain of our FCN^{SDEA} method over FCN is 1.1 (= 19.2 - 18.1), leading to a clear segmentation accuracy improvement of FCN^{SDEA}.

Besides demonstrating the accuracy improvement in terms of "mIoU" to help the readers to visualize the accuracy improvement by using FCN^{SDEA}, Fig. 3 depicts the perceptual effects of our SDEA approach. In Fig. 3, we observe that by using our FCN^{SDEA} method, the perceptual effects for the segmented road, sky, and building have been much improved relative to the FCN method.

To demonstrate the perceptual effect merit of the proposed SDEA approach for the two sources "leaves" and "manhole cover" as shown in the segmented roads of Fig. 3(b) and Fig. 3(d), our FCN^{SDEA} method justifies the IoU gains, 21.1 (= 86.8 - 65.7) and 25.2 (= 95.9 - 70.7), respectively, over the FCN method whose segmented roads are shown in Fig. 3(a) and Fig. 3(c). From the perceptual effects of the segmented sky and building, as shown in Fig. 3(f) and Fig. 3(h), our FCN^{SDEA} method justifies the IoU gains, 13.5 (= 95.4 - 81.9) and 36.0 (= 71.2 - 35.2), over the FCN method whose segmented results are shown in Fig. 3(e) and Fig. 3(g), respectively.

For fairness, based on the same testing dataset with 580 images [8], the average execution times for one testing image on the baseline model FCN and our model FCN^{SDEA} are reported. Because the CNN configurations of FCN and FCN^{SDEA} are the same and the only difference is the trained weights in the two models, for one testing image, the average execution times required by both models are the same and it takes 0.06 seconds.

B. OBJECT SEGMENTATION ACCURACY IMPROVEMENT MERIT OF SEGNET-BASIC^{SDEA}

In the second set of experiments, Table 4 tabulates the IoU comparison between the proposed SegNet-basic^{SDEA} method and the SegNet-basic method. In the last column of Table 4, the mIoU gain of our SegNet-basic^{SDEA} method over SegNet-basic is 3.1 (= 24.7 - 21.6), indicating a clear average IoU improvement by using our SDEA approach. In addition, as shown in Fig. 4, we observe that by using our SegNet-basic^{SDEA} method, the perceptual effects of the segmented road, sky, and building justify the related IoU improvements.

Based on the same testing dataset [8], the average execution times for one testing image on the baseline model SegNet-basic and our model SegNet-basic^{SDEA} are reported. Because the CNN configurations of the two models are the same and the only difference is the trained weights in the two models, for one testing image, the average execution times required by both models are the same and it takes 0.059 seconds.

C. OBJECT SEGMENTATION ACCURACY IMPROVEMENT MERIT OF ADAPTSEGNET^{SDEA}

In Table 5, the mIoU gain of our AdaptSegNet^{SDEA} method over AdaptSegNet is 1.5 (= 34.4 - 32.9), and it indicates a clear average IoU improvement by our SDEA approach. In Fig. 5, we observe that by using our AdaptSegNet^{SDEA} method, the perceptual effects of the segmented road, sky, and building justify the related IoU improvements.

Based on the same testing dataset, the average execution times for one testing image on the baseline model AdaptSegNet and our model AdaptSegNet^{SDEA} are the same because the configurations of the two CNN models are the same. Experimental results demonstrated that for one testing image, the average execution times required by both models are the same and it takes 0.036 seconds.

D. OBJECT SEGMENTATION ACCURACY IMPROVEMENT MERIT OF GATED-ADAPTSEGNET^{SDEA}

In Table 6, the mIoU gain of our Gated-AdaptSegNet^{SDEA} method over Gated-AdaptSegNet is 1.7 (= 36.4 - 34.7), indicating a substantial average IoU improvement by using our SDEA approach. In addition, Fig. 6 illustrates the perceptual effects of the segmented road, sky, and building by using our SDEA approach relative to Gated-AdaptSegNet.

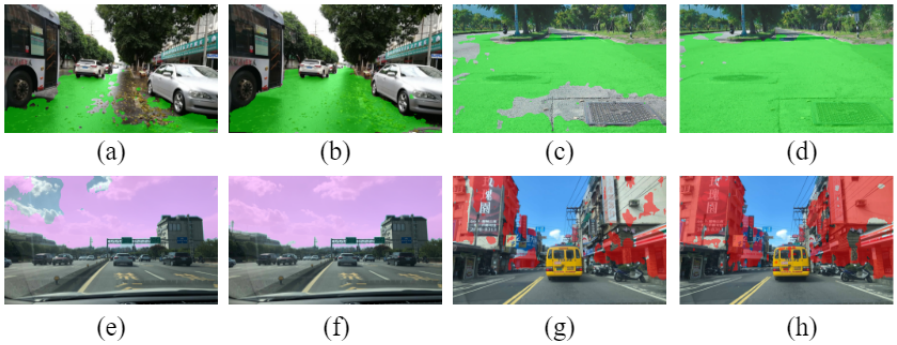


FIGURE 3. Perceptual effect merit of our FCN^{SDEA} method. (a) The first segmented road example with IoU = 65.7 by using FCN. (b) The first segmented road example with IoU = 86.8 by using our FCN^{SDEA} method. (c) The second segmented road example with IoU = 70.7 by using FCN. (d) The second road example with IoU = 95.9 by using our FCN^{SDEA} method. (e) The segmented sky example with IoU = 81.9 by using FCN. (f) The segmented sky example with IoU = 95.4 by using our FCN^{SDEA} method. (g) The segmented building example with IoU = 35.2 by using FCN. (h) The segmented building example with IoU = 71.2 by using our FCN^{SDEA} method.

TABLE 4. The mIoU improvement merit of the proposed SDEA approach relative to the SegNet-basic method [1].

	road	sidewalk	building	wall	fence	pole	light	sign	veg	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
SegNet-basic	68.6	17.3	54.1	24.1	1.4	26.3	3.8	6.8	62.1	7.0	59.2	6.7	11.3	51.0	4.7	1.3	0.0	0.2	4.1	21.6
SegNet-basic^{SDEA}	72.3	26.3	57.4	35.1	3.7	24.6	3.6	11.7	63.8	23.9	61.2	8.3	10.7	54.6	5.2	2.7	0.0	0.1	3.4	24.7

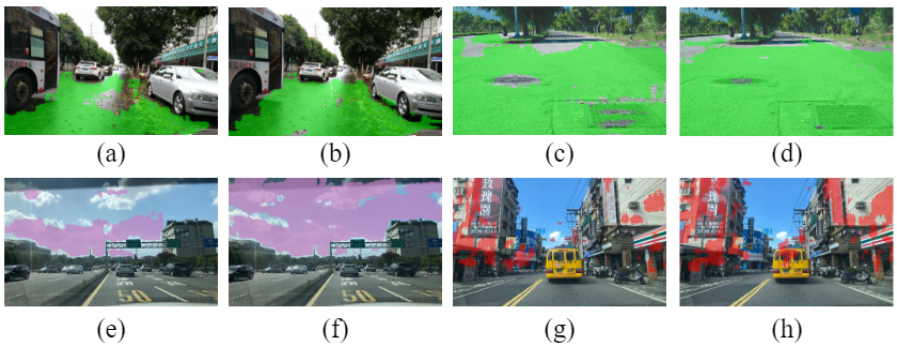


FIGURE 4. Perceptual effect merit of our SegNet-basic^{SDEA} method. (a) The first segmented road example with IoU = 75.0 by using SegNet-basic. (b) The first segmented road example with IoU = 80.1 by using our SegNet-basic^{SDEA} method. (c) The second segmented road example with IoU = 85.4 by using SegNet-basic. (d) The second road example with IoU = 90.4 by using our SegNet-basic^{SDEA} method. (e) The segmented sky example with IoU = 29.1 by using SegNet-basic. (f) The segmented sky example with IoU = 73.5 by using our SegNet-basic^{SDEA} method. (g) The segmented building example with IoU = 18.8 by using SegNet-basic. (h) The segmented building example with IoU = 28.6 by using our SegNet-basic^{SDEA} method.

TABLE 5. The mIoU improvement merit of the proposed SDEA approach relative to the AdaptSegNet method [27].

	road	sidewalk	building	wall	fence	pole	light	sign	veg	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
AdaptSegNet	80.7	16.6	79.1	12.7	15.7	24.5	17.8	17.7	76.9	11.3	78.5	38.5	8.6	79.8	22.2	26.7	0.7	17.8	0.1	32.9
AdaptSegNet^{SDEA}	83.0	22.3	79.3	17.5	17.0	26.8	15.7	12.4	77.3	11.3	78.2	41.0	15.3	80.2	21.9	32.4	0.1	20.6	1.3	34.4

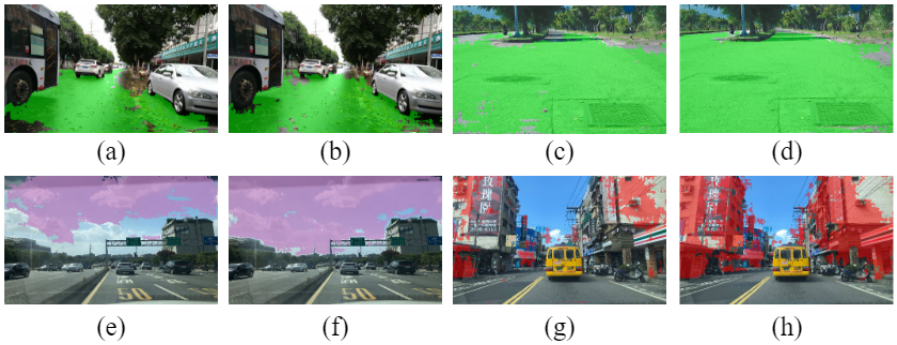


FIGURE 5. Perceptual effect merit of our AdaptSegNet^{SDEA} method. (a) The first segmented road example with IoU = 78.2 by using AdaptSegNet. (b) The first segmented road example with IoU = 80.7 by using our AdaptSegNet^{SDEA} method. (c) The second segmented road example with IoU = 91.6 by using AdaptSegNet. (d) The second road example with IoU = 94.9 by using our AdaptSegNet^{SDEA} method. (e) The segmented sky example with IoU = 64.8 by using AdaptSegNet. (f) The segmented sky example with IoU = 88.8 by using our AdaptSegNet^{SDEA} method. (g) The segmented building example with IoU = 20.7 by using AdaptSegNet. (h) The segmented building example with IoU = 65.3 by using our AdaptSegNet^{SDEA} method.

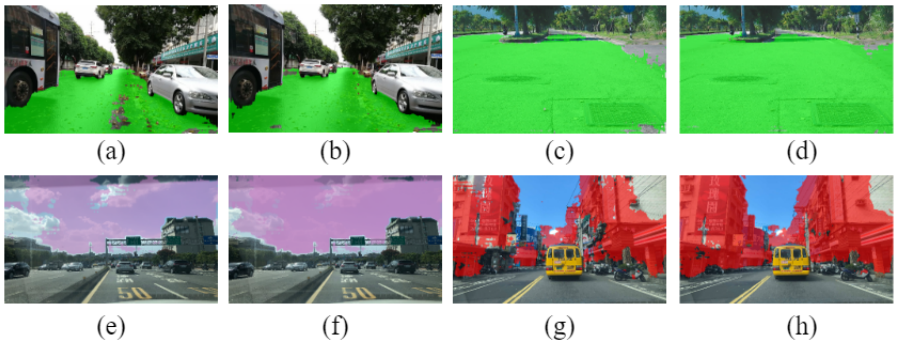


FIGURE 6. Perceptual effect merit of our Gated-AdaptSegNet^{SDEA} method. (a) The first segmented road example with IoU = 89.7 by using Gated-AdaptSegNet. (b) The first segmented road example with IoU = 92.4 by using our Gated-AdaptSegNet^{SDEA} method. (c) The second segmented road example with IoU = 92.8 by using Gated-AdaptSegNet. (d) The second road example with IoU = 96.9 by using our Gated-AdaptSegNet^{SDEA} method. (e) The segmented sky example with IoU = 74.9 by using Gated-AdaptSegNet. (f) The segmented sky example with IoU = 88.8 by using our Gated-AdaptSegNet^{SDEA} method. (g) The segmented building example with IoU = 64.7 by using Gated-AdaptSegNet. (h) The segmented building example with IoU = 80.4 by using our Gated-AdaptSegNet^{SDEA} method.

TABLE 6. The mIoU improvement merit of the proposed SDEA approach relative to the Gated-AdaptSegNet method [17].

	road	sidewalk	building	wall	fence	pole	light	sign	veg	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
Gated-AdaptSegNet	85.8	17.6	81.9	24.1	15.3	27.5	20.0	14.2	77.5	18.3	73.0	29.4	15.1	79.6	21.9	20.7	0.8	25.2	12.2	34.7
Gated-AdaptSegNet^{SDEA}	89.0	20.3	79.8	23.7	18.0	28.2	16.2	12.8	78.2	18.2	74.8	29.4	15.7	84.1	30.2	35.5	0.1	18.5	18.4	36.4

Based on the same testing dataset, for one testing image, the execution times required by the baseline model Gated-AdaptSegNet and our model Gated-AdaptSegNet^{SDEA} are the same and it takes 0.042 seconds.

IV. CONCLUSION

We have presented the proposed novel and effective SDEA approach to enhance the accuracy of the CNN-based object segmentation methods on FCN, SegNet-basic, AdaptSegNet, and Gated-AdaptSegNet. In particular, in the proposed fast source-pasting technique, the labelled pixel-annotations covered by these sources can inherit the original pixel-annotations from the sub-image of the selected "Camvid" image, and the labelled pixel-annotations covered by the other parts can also herit the original pixel-annotations in the selected GTA5 image. The comprehensive experimental results have justified the segmentation accuracy improvement merit and the perceptual effect of our SDEA approach relative to the four CNN-based object segmentation methods on FCN, SegNet-basic, AdaptSegNet, and Gated-AdaptSegNet.

Our first future work is to extend our SDEA approach to cover more sources for further improving existing CNN-based object segmentation methods. In addition, our SDEA approach will be considered to apply to the spatio-

temporal graph convolutional network-based traffic forecasting method [31] which has been successfully used in the public bike sharing program [30]. Our second future work is to deploy the time-varying communication time delay issue [6], [28] into the proposed SDEA- and CNN-based object segmentation method to achieve higher segmentation accuracy and real-time demand in ADAS applications.

ACKNOWLEDGMENTS

We appreciate the help of Associate Editor Dr. B. Wang and the three anonymous reviewers for their valuable comments to improve the manuscript. We also appreciate the proofreading help of Ms. C. Harrington.

REFERENCES

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 39, no. 12, pp. 2481-2495, Dec. 2017.

[2] 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, Jan. 2009.

[3] C. Chang, J. Zhao, and L. Itti, "DeepVP: deep learning for vanishing point detection on 1 million street view images," *IEEE International Conference on Robotics and Automation,* Brisbane, Australia, 2018, pp. 4496-4503.

[4] Y. Chen, W. Li, and L. Van Gool, "ROAD: reality oriented adaptation for semantic segmentation of urban scenes," *IEEE Conference on Computer Vision and Pattern Recognition,* Salt Lake City, UT, USA, 2018, pp. 7892-7901.

[5] 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," *IEEE Conference on Computer Vision and Pattern Recognition,* Las Vegas, NV, USA, 2016, pp. 3213-3223.

[6] C. Deng, M. Er, G. Yang, and N. Wang, "Event-triggered consensus of linear multiagent systems with time-varying communication delays," *IEEE Transactions on Cybernetics,* Early Access, Aug. 2019.

[7] D. Ding, C. Lee, and K. Lee, "An adaptive road ROI determination algorithm for lane detection," *IEEE International Conference of IEEE Region 10,* Xi'an, China, 2013, pp. 1-4.

[8] Execution code. Accessed: 26 Mar. 2020. [Online]. Available: ftp://140.118.175.164/Images.

[9] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research,* vol. 32, no. 11, pp. 1231–1237, Sep. 2013.

[10] R. Girshick, "Fast R-CNN," *IEEE International Conference on Computer Vision,* Santiago, Chile, 2015, pp. 1440-1448.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *IEEE Conference on Computer Vision and Pattern Recognition,* Columbus, OH, USA, 2014, pp. 580-587.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE International Conference on Computer Vision,* Venice, Italy, 2017, pp. 2980-2988.

[13] H. Hwang, G. Yoon, and S. Yoon, "Optimized clustering scheme-based robust vanishing point detection," *IEEE Transactions on Intelligent Transportation Systems,* vol. 21, no. 1, pp. 199-208, Jan. 2020.

[14] C. Lee and J. Moon, "Robust lane detection and tracking for real-time applications," *IEEE Transactions on Intelligent Transportation Systems,* vol. 19, no. 12, pp. 4043-4048, Dec. 2018.

[15] Y. Li, M. Liu, X. Li, M. Yang, and J. Kautz, "A closed-form solution to photorealistic image stylization", *European Conference on Computer Vision,* Munich, Germany, 2018, pp. 453-468.

[16] J. Li and Z. Wang, "Real-time traffic sign recognition based on efficient CNNs in the wild," *IEEE Transactions on Intelligent Transportation Systems,* in early access, 2020.

[17] Y. Lin, D. Tan, W. Cheng, and K. Hua, "Adapting semantic segmentation of urban scenes via mask-aware gated discriminator," *IEEE International Conference on Multimedia and Expo,* Shanghai, China, 2019, pp. 218-223.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Conference on Computer Vision and Pattern Recognition,* Boston, MA, USA, 2015, pp. 3431-3440.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 39, no. 6, pp. 1137-1149, Jun. 2017.

[20] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," *European Conference on Computer Vision,* Amsterdam, Netherlands, 2016, pp. 102-118.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *International Conference on Medical Image Computing and Computer Assisted Intervention,* Munich, Germany, 2015, pp. 234–241.

[22] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," *IEEE Conference on Computer Vision and Pattern Recognition,* Las Vegas, NV, USA, 2016, pp. 3234-3243.

[23] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: a database and web-based tool for image annotation," *International Journal of Computer Vision,* vol. 77, no. 1-3, pp. 157-173, May 2008.

[24] F. S. Saleh, M. S. Aliakbarian, M. Salzmann, L. Petersson, and J. M. Alvarez, "Effective use of synthetic data for urban scene semantic segmentation," *European Conference on Computer Vision,* Munich, Germany, 2018, pp. 86-103.

[25] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 39, no. 4, pp. 640-651, Apr. 2017.

[26] W. Song, Y. Yang, M. Fu, Y. Li, and M. Wang, "Lane detection and classification for forward collision warning system based on stereo vision," *IEEE Sensors Journal,* vol. 18, no. 12, pp. 5151-5163, Jun. 2018.

[27] Y. Tsai, W. Hung, S. Schulter, K. Sohn, M. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," *IEEE Conference on Computer Vision and Pattern Recognition,* Salt Lake City, UT, USA, 2018, pp. 7472-7481.

[28] B. Wang, W. Chen, B. Zhang, and Y. Zhao, "Regulation cooperative control for heterogeneous uncertain chaotic systems with time delay: A synchronization errors estimation framework", *Automatica,* vol. 108, Oct. 2019, Art. no. 108486.

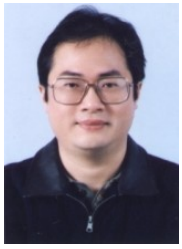
[29] C. Wu, L. Wang, and K. Wang, "Ultra-low complexity block-based lane detection and departure warning system," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 29, no. 2, pp. 582-593, Feb. 2019.

[30] G. Xiao, R. Wang, C. Zhang, and A. Ni, "Demand prediction for a public bike sharing program based on spatio-temporal graph convolutional networks," *Multimedia Tools and Applications,* in press, 2020.

[31] Yu B, Yin H, and Zhu Z "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," arXiv:1709.04875, 2017.

[32] Y. Zhang, P. David, H. Foroosh, and B. Gong, "A curriculum domain adaptation approach to the semantic segmentation of urban scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 42, no. 8, pp. 1823-1841, Aug. 2020.

[33] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, "Fully convolutional adaptation networks for semantic segmentation," *IEEE Conference on Computer Vision and Pattern Recognition,* Salt Lake City, UT, USA, 2018, pp. 6810-6818.



KUO-LIANG CHUNG (SM01) received his B.S., M.S., and Ph.D. degrees from National Taiwan University, Taipei, Taiwan, in 1982, 1984, and 1990, respectively. He has been one Chair Professor of the Department of Computer Science and Information Engineering at National Taiwan University of Science and Technology, Taipei, Taiwan since 2009. He was the recipient of the Distinguished Research Award (2004-2007; 2019-2022) and Distinguished Research Project Award (2009-2012) from the Ministry of Science and Technology of Taiwan. In 2020, he received the K. T. Li Fellow Award from the Institute of Information Computing Machinery, Taiwan. He has been the Associate Editor of the Journal of Visual Communication and Image Representation since 2011. His research interests include machine learning, image processing, and video compression.

YA-YUN CHENG received her B.S. degree in Computer Science and Information Engineering from the Tamkang University, New Taipei City, Taiwan, in 2017. She received her M.S. degree in Computer Science and Information Engineering at the National Taiwan University of Science and Technology, Taipei, in 2020. Her research interests include machine learning and object segmentation.