IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 September 2023, accepted 28 October 2023, date of publication 1 November 2023, date of current version 6 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3329302

RESEARCH ARTICLE

Context-Aware Region-Dependent Scale Proposals for Scale-Optimized Object Detection Using Super-Resolution

KAZUTOSHI AKITA AND NORIMICHI UKITA, (Member, IEEE)
Toyota Technological Institute, Nagoya 468-8511, Japan

Corresponding author: Kazutoshi Akita (sd21501@toyota-ti.ac.jp)

This work was supported in part by the Japan Society of the Promotion of Science (JSPS) KAKENHI under Grant 19K12129 and Grant 22H03618.

ABSTRACT Image scaling techniques such as Super-Resolution (SR) are useful for object detection, especially for detecting small objects. However, we found that scaling by an inappropriate factor tends to induce false-positive detections. This paper presents a Region-Dependent Scale-Proposal (RDSP) network that estimates the appropriate scale factors for each image region depending on its contextual information. In our detection framework, images are appropriately scaled by SR according to the estimations of the RDSP network, and fed into the scale-specific object detectors. While previous works have proposed models for scale proposal, our RDSP extracts regions where objects could potentially exist based on scene structure, regardless of whether actual objects are present, because small objects are often too small to determine their presence accurately. Additionally, while existing approaches have fused object detection and SR in an end-to-end manner, scale proposals for SR are not provided or are performed independently. Qualitative and quantitative experiments show that our RDSP network provides appropriate SR scales and improve detection accuracy on highly challenging dataset, captured by real car-mounted cameras with size-varied objects, including extremely small objects. Our code is available at https://github.com/kakitamedia/RDSP.

INDEX TERMS Object detection, object scales, region-dependent scale proposals, super-resolution, tiny object detection.

I. INTRODUCTION

Object detection is one of the most important computer vision tasks. One of the challenges in object detection lies in the variation in object sizes, and many previous works addressed this size-varied object detection task.

A standard anchor-based method detects size-varied objects from a wide variety of scaled anchor boxes that are densely distributed in an image (e.g., SSD [2], RetinaNet [3], and Faster R-CNN [4]). However, such a huge number of anchor boxes that cover a variety of locations and scales not only degrade computational efficiency but also make learning difficult.

In contrast to such a scale-invariant object detector, it is known that object size-varied object detection can be improved by scale-specific object detectors [5], [6] with image scalings. In SNIP [5], an image is scaled to various sizes by pre-determined upscale factors with image interpolation, and all the scaled images are fed into its corresponding scale-specific detector. However, it has been pointed out that image interpolation degrades the performance of small object detection [1].

To detect such small objects, detection methods using Super-Resolution (SR) have been proposed. For example, face detection [7] and generic object detection [1], [8]. In [7], [8], SR is applied to regions extracted by RPN. However, in cases where objects are extremely small, RPN may not perform effectively. Therefore, the entire detection process, including RPN, should performed on SR images. In [1], SR is

The associate editor coordinating the review of this manuscript and approving it for publication was Felix Albu.

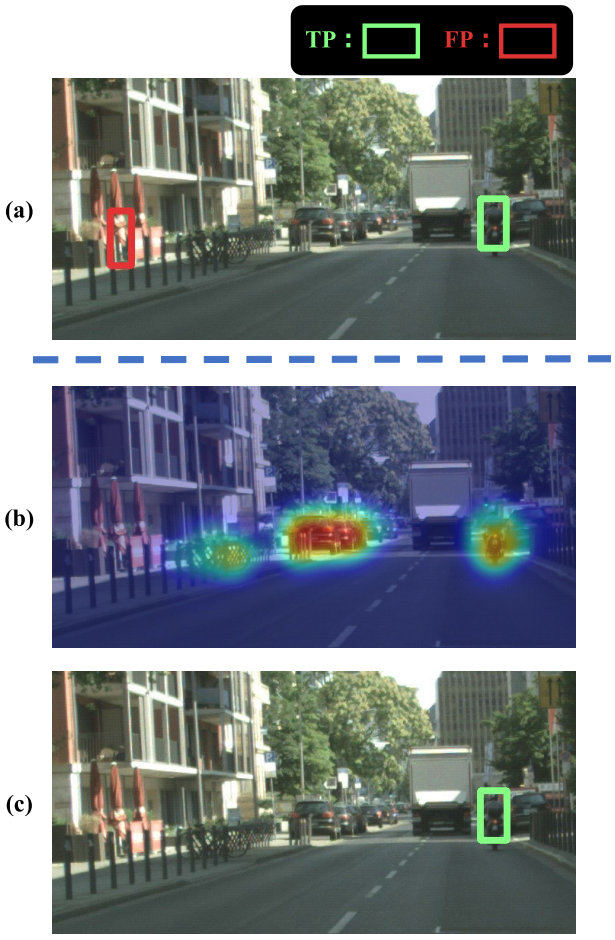

FIGURE 1. Examples of our object-scale proposals for scale-specific object detection with super-resolution. (a) Detection result by TDSR [1] with factor 2. False-positive detection can be seen on the left of the image. (b) SR-scale proposals by our Region-Dependent Scale-Proposal (RDSP) network. The regions appropriate for upscaling by super-resolution with factor 2 are predicted. (c) Detection result of our proposed detection pipeline. False-positive detection is suppressed based on (b).

FIGURE 2. Recall and precision of standard one-stage detectors with different settings. Naive application of SR improves the recall by achieving the detection of small objects, but harms the precision because of false-positives. Our method improved recall without harming precision.

applied before the entire detection process, but they conduct experiments only on manually downsized images. This means the effectiveness of SR for size-varied object detection is not evaluated.

Therefore, we apply SR directly to real data, which contains objects of various sizes, including extremely small objects. However, we found that application of SR to detectors tends to detect false-positives (as shown in Figure 1 (a) and Figure 2) when object regions are rescaled by inappropriate scaling factors, while these inappropriately-rescaled regions are not false-detected in the original-scale
image.

This paper presents how to estimate the appropriate scale of each image region. Our contributions in this paper are as follows:

- We apply SR to the size-varied object detection tasks, while the previous SR-based approach applies SR to manually downscaled images.
- Our proposed method addresses false-positive detections in SR-based object detection. In our method, an appropriate SR scale is predicted at each image region depending on the scene structure. Based on this prediction, the false-positive detections caused by inappropriate scaling are suppressed.
- The aforementioned appropriate scale prediction is achieved by our proposed network called a Region-Dependent Scale-Proposal (RDSP) network. While some previous works have proposed models for scale proposal, our RDSP extracts regions where objects could potentially exist based on scene structure, regardless of whether actual objects are present, because small objects are often too small to determine their presence accurately.
- In RDSP, the global scene structure and local appearance features are utilized in implicit and explicit manners. While RDSP is presented in our early work [9], it is extended with the positional and global structure embeddings in this paper.
- This paper explores how to effectively end-to-end train the network consisting of RDSP, SR, and detection sub-networks. Since the full network is huge, the ill-considered combination of complex loss functions makes it difficult to train the network due to conflict between the losses. We found the best combination of the losses for end-to-end learning of the full network.
- Our method can be applied to any differentiable object detector. This applicability is demonstrated in our experiments shown in this paper. Various object detectors are integrated with RDSP in order to improve the performance of object detection.

- To validate the effectiveness of the proposed method, we utilize datasets captured by car-mounted cameras, which have severe object size variations, including extremely small objects.

While the aforementioned first and second contribution has been presented in our early version [9], the remaining contributions are the novel contributions presented in this paper.

II. RELATED WORK

In this section, we introduce SR using deep convolutional networks (in Sec. II-A) and object detection using SR and/or scene-object relationships (in Secs. II-B and II-C).

A. DEEP SUPER-RESOLUTION

As with many computer vision technologies, SR has been improved with convolutional networks (e.g., DBPN [10], WDST [11], SRFlow [12], PAMS [13], CARN [14], LatticeNet [15], SRNTT [16], SPSR [17]). Recent approaches with downscaling kernel representations [18], [19], [20], [21], [22], [23] improve the SR performance and its applicability to real-world images degraded by a variety of blur kernels. Since SR is one of the hot topics in computer vision as demonstrated in public challenges [24], [25], [26]. Furthermore, single-image SR is extended to video SR [27], [28], [29], [30], [31] and joint space-time video SR [32], [33], [34], [35] for more variety of applications.

However, all of these SR methods are designed to improve the image quality for human perception, which is evaluated by PSNR and other image-quality metrics. While our proposed method can utilize any of these SR methods, the goal of our work is to explore SR methods applicable to machine perception, such as object detection. To this end, in previous work, SR is combined with an object detector in tiny face detection [7] and tiny generic object detection [1], [8]. This paper proposes automatic region-dependent scale proposals for SR, in addition to image upscaling using SR.

B. OBJECT SCALE PROPOSAL

To address size-varied object detection challenges, numerous strategies have been proposed. For example, image pyramid-based [5], [6], [36], and feature pyramid-based [2], [3], [37] strategies are common practice. However, these methods are unsuitable for scenarios with extremely large variations in object sizes, as they require constructing very large pyramids. Therefore, previous works [38], [39] proposed that explicitly predict object sizes in advance and perform detection based on these predictions. However, since these approaches rely on the appearance of objects in the image to estimate their scale, estimation can fail when dealing with extremely small objects. Therefore, scale estimation methods that do not rely on object appearance are needed.

Scene structure also plays a role in determining the scale of objects in an image; for instance, a sidewalk near the vanishing point may contain a small person. Hence, some studies explicitly utilize this kind of scene structure. For example, scene-specific [40], [41], perspective-aware [42], depth-aware [43], 3D geometry-aware [44] object detectors are proposed.

In this study, we construct the RDSP network that incorporates scene structures and propose its training method that enables scale estimation independent of object appearance.

C. SCALE-DEPENDENT OBJECT DETECTION

Beyond object scale recognition [42], [43], [44], [45] mentioned in Sec. II-B, local image regions can be explicitly rescaled for more easy recognition (e.g., region-dependent object scaling for object detection [46], in particular for tiny object detection [47]). Tiny object detection can be further improved by incorporating SR and detection networks with end-to-end learning [1].

Our proposed method integrates scale-dependent detection with SR so that appropriate scales in different image local regions are estimated. Previous methods (i) estimate a region-independent scale histogram [46] or (ii) just employ off-the-shelf detectors for estimating region-dependent scale proposals [47]. On the other hand, our method has additional networks (RDSP) for proposing the probability maps, in which a pixel value in each pixel is higher if it is highly possible that any target object is observed in that pixel, for multiple SR scales. Each scale proposal is expressed as a heatmap image, as shown in Figure 1 (b).

III. REGION-DEPENDENT SR-SCALE PROPOSALS FOR SCALE-SPECIFIC OBJECT DETECTION

Figure 3 shows the overall pipeline of our proposed method. Our detection pipeline consists of three independent branches and one independent detector, and each branch is composed of the following processes. Note that each ''term'' enclosed in double quotation marks appears in Figure 3 and Figure 4.

1) **SR networks:** The SR network is prepared for each scaling factor.
2) **RDSP network:** The RDSP network is designed to estimate the ''Scale-proposal heatmaps'' where the appropriateness of each SR scaling factor (e.g., factors of 1, 2, and 4) is given at each pixel. The examples of the scale proposal heatmap are shown in Figure 1 (b).
3) **Detection network:** Each of the generated ''Masked SR images'' is fed into its corresponding-scale object detection network (i.e., ''Detector'' in Figure 3). Detected bounding boxes are superimposed onto a single image, where overlapping bounding boxes are merged by a standard non-maximum suppression scheme.

Our RDSP network suppresses false positives due to inappropriate scaling by proposing an appropriate scaling factor for each image region. In what follows, we explain our proposed ''RDSP network'' in detail in Section III-A. After how to train the network consisting of three networks (RDSP network, SR, and detection) is introduced in Section III-B,

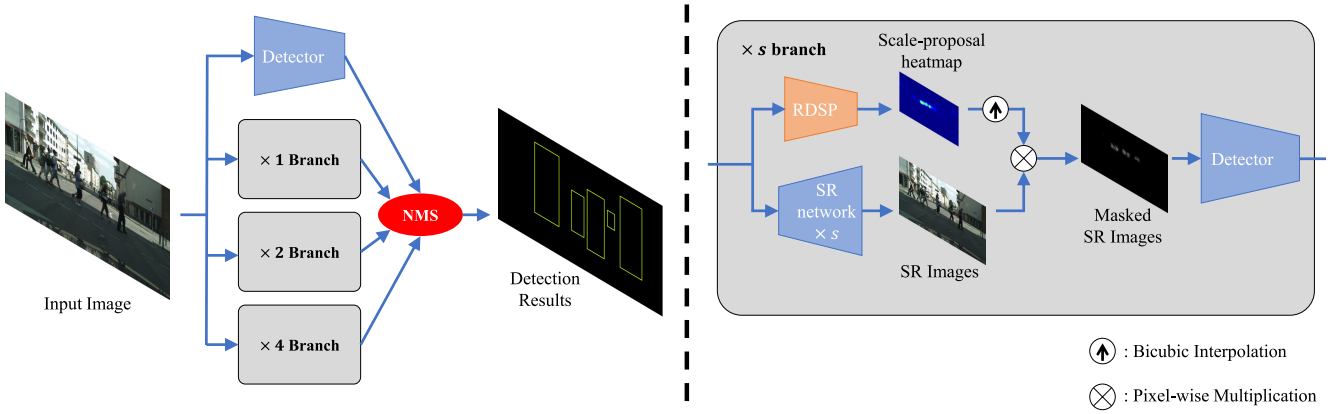

FIGURE 3. Overall pipeline of the proposed method. Any differentiable network can be used for the SR network and detector. The detailed structure of RDSP networks is shown in Figure 4.

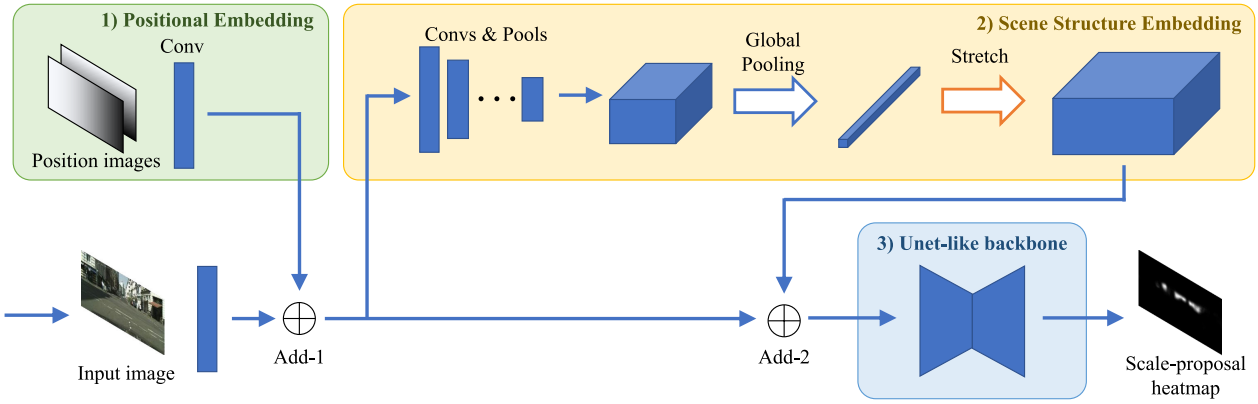

FIGURE 4. RDSP architecture. ⊕ denotes a pixel-wise add operation.

Section III-C describes how to create the ground-truth for the training of RDSP network.

A. REGION-DEPENDENT SR-SCALE PROPOSALS

The RDSP network is required to roughly but robustly detect regions in each of which there might be any object. This region is called a Possible Object Region (POR). In particular, even PORs of tiny objects must also be detected by the RDSP network. Such PORs of tiny objects are upscaled using SR by large scaling factors (e.g., the factor of 4). However, this scheme seems to be a chicken-and-egg problem because the RDSP network must detect PORs of tiny objects to support the following object detection network. Therefore, the goal of the RDSP network is not to precisely detect objects without excess or deficiency but to roughly detect PORs with no false negatives. While the POR is similar to a general region proposal for object detection, RDSP also estimates the appropriate scaling factor of each POR for improving the performance of scale-specific object detection. One more difference between the region proposal and the POR is that a set of pixelwise probability values is provided in a heatmap image for our POR representation, while the region proposal is a bounding box corresponding to each object region.

Since POR estimations for tiny objects are difficult, we need additional cues as well as the appearance information of each region of interest. To encode additional cues, the proposed RDSP architecture has the following three modules (also shown in Figure 4):

1) **"Positional embedding":** If similar scene structures are observed in different images, objects in each class are observed to be similar sizes in specific pixel coordinates in these images. For example, in the case of an in-vehicle camera, small objects are located near the vanishing point, i.e., around the center of the image. In order to utilize such a constraint implicitly, we employ a positional embedding scheme. For this embedding, we use a pair of "position images" consisting of x and y channels in which the value of each pixel is x and y image coordinate, respectively. These position images are fed into a 1×1 conv layer. The output features of this conv layer are pixelwise added to image features extracted from the input image (as indicated by "Add-1" in the figure).

2) **"Scene structure embedding":** If the structure of a scene is neglected, ridiculous detections might be found. For example, the misclassification of a boat in the sea as a car is a typical one [48]. To suppress such misclassification, methods using the structure of the entire scene have been proposed. (e.g., based on object-scene relationships [48] and attention [49]). In our work, scene structure influences the image features by adding the pooled global features to the image features pixel-wisely (as indicated by "Add-2" in the figure).
3) **"UNet-like backbone":** The effectiveness of the context around a region of interest is validated for tiny object detection such as face detection [7], human-body key-point detection [50], and semantic segmentation [51]. The proposed method employs a semantic segmentation network, U-Net [52], for estimating the scale proposals by jointly evaluating local and global appearance features. The U-Net architecture [52] has symmetric expanding paths for precise localization and contracting paths for capturing contexts. These two types of paths are designed to be symmetric and connected in order to propagate a large number of feature channels from the contracting paths to the expanding paths for utilizing the local and global contexts.

The "Add" operation described above refers to element-wise addition defined by the following equation.

$$F'_{x,y,c} = F_{x,y,c} + E_{x,y,c} \tag{1}$$

where F' denote the post-embedded feature, while F denotes the image feature and E denotes the embedding feature. x, y, and c denote positions on each feature map.

The "stretch" operation in Figure 4 involves replicating a 1-dimensional vector V in both the height and width dimensions to create a 3-dimensional feature map. The "stretch"ed feature map M is expressed as the following formula:

$$M_{x,y,c} = V_c \tag{2}$$

Through this "stretch" operation, the pooled 1-dimensional feature vector is expanded to the same size as the image feature map.

The RDSP network represents the estimated PORs as a set of heatmap images, each of which corresponds to each scaling factor, as indicated by the "Scale-proposal heatmap" in Figure 4. Since the last layer of the UNet-like backbone is a Sigmoid activation layer, each pixel value in the scale-proposal heatmap is normalized between 0 and 1, which means the less probable and the most probable to be a POR, respectively. As shown in Figure 3, our proposed network has multiple RDSP networks, each of which corresponds to each scaling factor.

While our proposed RDSP is similar to the general Region Proposal Network (RPN), it differs in that RDSP estimates the probability of an object's presence based solely on context, regardless of whether the object actually exists there. In contrast to RPN, which is trained based on the ground truth bounding boxes, our RPSP is trained to inherently extract regions where objects could potentially exist by our proposed training algorithm, as detailed in the next section.

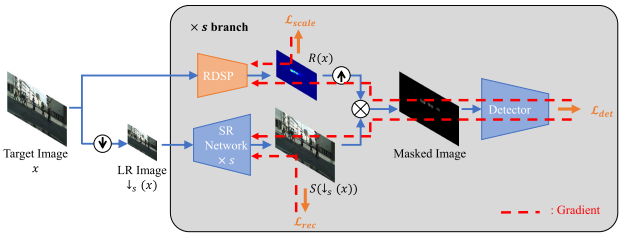

FIGURE 5. The overview of the training of our proposed method.

B. END-TO-END TRAINING FOR SR AND RDSP

In our proposed detection framework, we train each branch (as illustrated on the right in Figure 3) independently. An overview of the training for each branch is shown in Figure 5.

As well as a detection network, any differentiable SR network can be utilized for our joint SR and detection network. Prior to end-to-end training of the full network consisting of these two networks, we assume that each of them is pre-trained in accordance with the training process of each network for better training of the full network. In what follows, the end-to-end training scheme following these pre-training processes is described.

As with the basic training process of an SR network, in our end-to-end training scheme, the SR network is trained with the following reconstruction loss expressed by the mean absolute error (MAE):

$$L_{rec}(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} |x_i - \hat{x}_i|, \tag{3}$$

where x and $\hat{x}$ denote the ground-truth HR image and its SR image, respectively. N is the number of pixels in each image.

In contrast to general SR training using L_{rec}, we train SR networks also with a detection loss (denoted by L_{det}) for optimizing SR for object detection. Let $\downarrow_s (\cdot)$ denote an image downscaling function by factor s, and the SR network is denoted by S. The detection network D takes the SR image masked by the heatmap estimated by the RDSP (denoted by R). The form of the output of D (i.e., detection results) and its ground truth (denoted by y) differs depending on the object detector. For example, in CenterNet [53], y consists of three kinds of multi-channel images. The pixel values of the first, second, and third images are (i) the confidence value of object center detection in each pixel, (ii) the width and height of the object bounding box in each pixel, and (iii) bounding-box displacements caused by the output stride; see the original paper [53] for details.

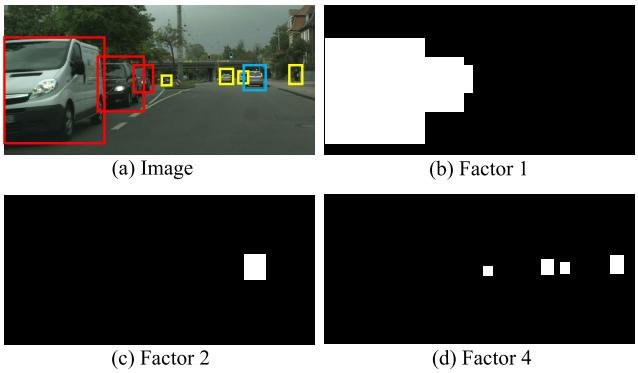

FIGURE 6. (a) Training image and ground-truth bounding boxes. Red, blue, and yellow bounding boxes are grouped into those of factors 1, 2, and 4, respectively. From this training image, ground-truth heatmap images for each factor are generated as shown in (b), (c), and (d), respectively. These heatmaps are blurred by Gaussian for robust detection.

With L_{rec} and L_{det}, the SR network in our full network is trained with the following compound loss function L_{SR}:

$$L_{SR}(x, y) = \lambda_{rec} L_{rec}(x, S(\downarrow_s (x))) + \lambda_{det} L_{det}(y, D(S(\downarrow_s (x)) \odot R(x))), \quad (4)$$

where λ_{rec} and λ_{det} are constants determining the relative weights of the reconstruction loss and the detection loss, respectively.

The RDSP network is trained with the following loss expressed by the binary cross entropy (BCE) using the ground-truth heatmap that is created from the detection ground-truth:

$$L_{scale}(p, \hat{p}) = -\frac{1}{N} \sum_{i=1}^{N} p_i \log \hat{p}_i \quad (5)$$

where p and $\hat{p}$ denote the ground-truth heatmap and the predicted heatmap, respectively. As described above, the RDSP network is required to roughly but robustly detect possible object regions. Therefore, the ground-truth heatmap p is designed to fulfill that requirement. The details of ground-truth heatmap p are described in Sec III-C.

We train the RDSP networks also with a detection loss for object detection. With L_{scale} and L_{det}, the RDSP network is trained with the following loss function L_{RDSP}:

$$L_{RDSP}(x, y, p) = \lambda_{scale} L_{scale}(p, R(x)) + \lambda_{det} L_{det}(y, D(S(\downarrow_s (x)) \odot R(x))), \quad (6)$$

where λ_{scale} are weighting parameter for L_{scale}.

In total, the loss function for the entire proposed framework is as follows:

$$L_{total} = L_{RDSP} + L_{SR} \quad (7)$$

C. SCALE-PROPOSAL GROUND-TRUTH FOR RDSP TRAINING

As described in Sec. III-B, the RDSP network is trained with the ground-truth heatmaps. Although the RDSP networks are required to estimate the regions that are suitable for the corresponding scale factor, it is difficult to learn to satisfy such requirements from end-to-end training with object detection alone. Therefore, to support the training, we create ground-truth heatmaps that satisfy the requirements. For producing the ground-truth data for this training, a standard training dataset for object detection is reprocessed as follows:

1) The bounding boxes of objects for detection are divided into height-dependent groups. In our experiments, the bounding boxes are divided into three groups, namely bounding boxes whose appropriate scaling factors are 1, 2, and 4. More specifically, in our experiments, the groups of scaling factors of 1, 2, and 4 include the following ranges depending on the height of the bounding box (denoted by h_b): (1) if $h_b \geq 64$, factor of 1, (2) if $32 \leq h_b < 64$, factor of 2, and (3) if $h_b < 32$, factor of 4. In the training images, the groups of factors 1, 2, and 4 have 7,156, 6,348, and 6,297 bounding boxes, respectively.
2) Each RDSP network produces a heatmap-like image in which higher values are given in pixels where any target object is likely to be observed. The ground-truth image of the heatmap for the factor of $S \in 1, 2, 4$ contains only the bounding boxes included in factor S's group, as illustrated in Figure 6. In each ground-truth image, all bounding boxes are filled by 1, while all other pixels are 0.
3) For robust detection, all bounding boxes filled by 1 are blurred by Gaussian. This blurred image is used as the ground truth of the output of the RDSP network for training.

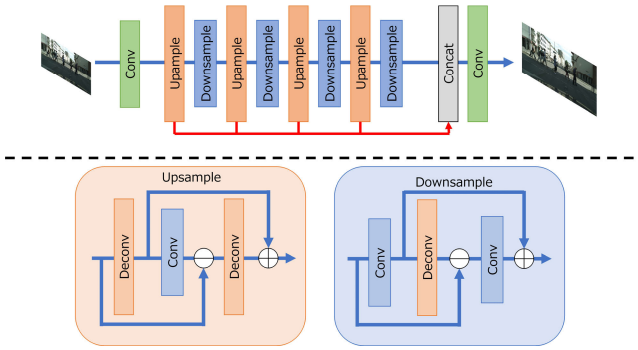

FIGURE 7. The architecture of DBPN [10] network.

IV. EXPERIMENTS

A. DATASET

We conducted experiments with the CityScapes dataset [54], which is a car-mounted camera dataset. Since this dataset is developed for evaluating instance segmentation methods, the annotations are pixelwise instance labels. From these pixelwise instance labels, bounding boxes for object detection were generated so that a rectangle circumscribing each instance is regarded as its bounding box. While 30 object

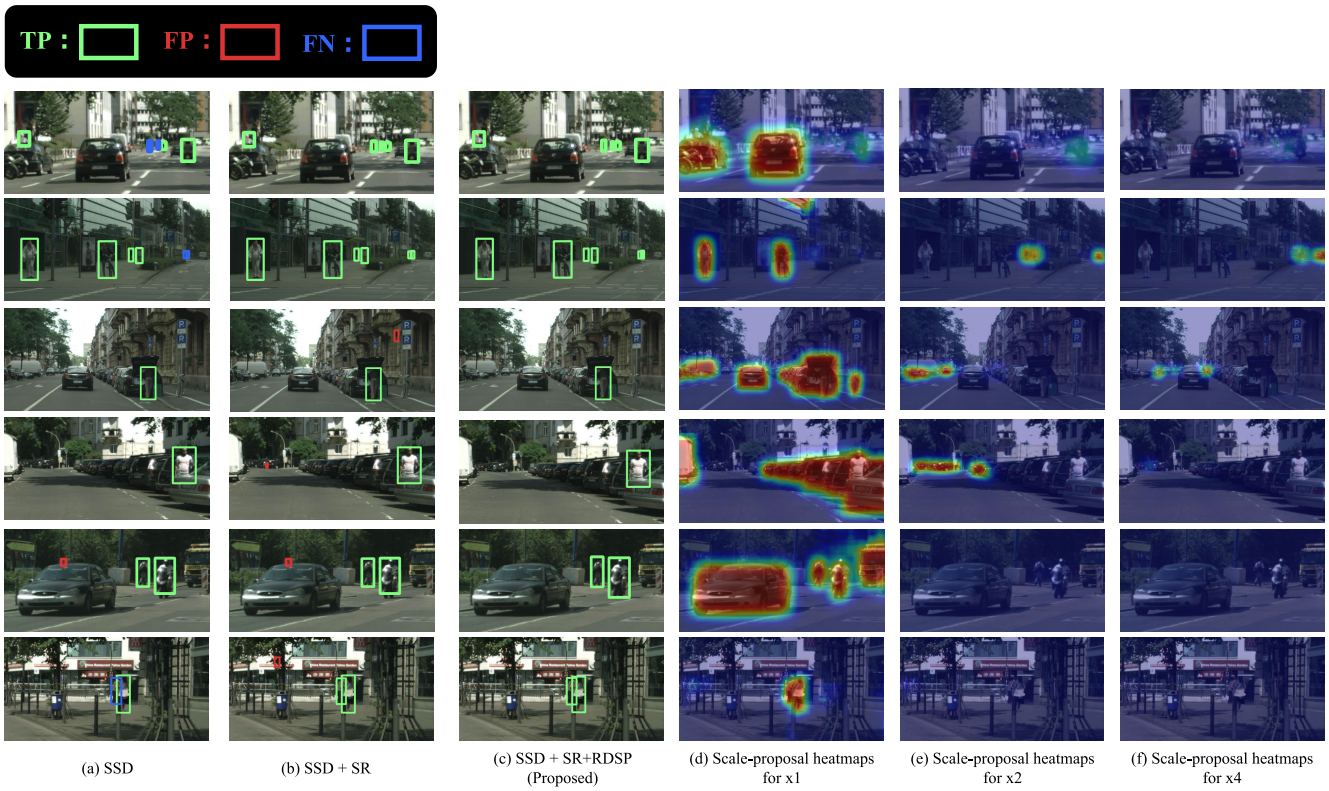

FIGURE 8. The successful cases on the CityScapes dataset with Super-Resolution and the proposed RDSP. The predicted bboxes with confidence greater than 50% are visualized. For better visualization, a part of the image is cropped.

classes are defined in the dataset, most of them are background objects such as "sky" "road," and "vegetation." In our experiments, only classes included in the human group (i.e., "person" and "rider") were used for the object detection task. The CityScapes dataset is officially split into training and test images. The number of training and test images containing human bounding boxes is 2,965 and 492, respectively. In all the training and test images, 19,801 and 3,975 human bounding boxes are included, respectively.

Additionally, we conducted experiments on the BDD100k dataset, which is a dataset collected from car-mounted cameras, like the CityScapes dataset. In contrast to the CityScapes dataset, the BDD100k dataset is primarily designed for object detection tasks, so we use the provided bounding boxes as-is. The BDD100k dataset is officially split into training and test images. The training set contains 92,369 human bounding boxes, while the test set has 13,426 human bounding boxes. The dataset consists of a total of 70,000 training images and 10,000 test images.

B. TRAINING DETAILS

The proposed method has three components: SR network, RDSP network, and object detection network. As an SR network, we use Deep Back-Projection Network (DBPN) [10], which achieves competitive results in SR challenges [24]. The architecture of DBPN is shown in Figure 7. As a detection network, we use the following five detectors of three types: (1) One-stage detector; SSD [2] and RetinaNet [3], (2) Two-stage detector; Faster R-CNN [4], (3) Anchor-free detector; FCOS [55], CenterNet [53].

First, these three components are pretrained independently by L_{rec}, L_{scale}, L_{det}, respectively. For this pretraining, we use Adam [56] optimizer with β=(0.9, 0.999), and the mini-batch size is 8. The learning rate is initialized to 1e-4 and multiplied by 1/10 at 300,000 and 450,000 iterations, while the total iterations are 500,000. As augmentations, we apply random flipping and random cropping to 512 $\times$ 512. The weights of SR networks are initialized from the author model, while the weights of the detection networks are initialized from coco-pretrained weights published in mmdetection [57] project. The RDSP models are trained from randomly-initialized weights.

After this pretraining, these networks are fine-tuned in an end-to-end manner with loss functions described in Sec III-B. For this training, we set the mini-batch size 6. The learning rate is initialized to 1e-4 and multiplied by 1/10 at 200,000 and 280,000 iterations, while the total iterations are 300,000. Other settings follow the pretraining.

C. RESULTS AND ANALYSIS

1) EFFECTS OF SR AND RDSP

We compare the results with and without SR and RDSP to validate their effectiveness. The qualitative results on

TABLE 1. Experimental results on the CityScapes dataset. We conducted comparisons across various object detectors (i.e., SSD, RetinaNet, Faster R-CNN, FCOS, and CenterNet) with and without SR and our RDSP. ✓denotes that the corresponding sub-network is integrated into the full network. Red and blue values indicate the best and the second-best scores in each detector, respectively.

Model		SR	RDSP	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
One-stage detector	SSD [2]			37.3	66.1	37.0	17.8	46.3	60.0
		✓		39.5	68.3	39.7	20.4	48.5	59.2
		✓	✓	40.3	70.8	40.3	22.0	49.0	59.8
	RetinaNet [3]			39.2	68.3	39.7	22.0	47.8	58.6
		✓		41.1	71.0	41.3	24.8	48.6	58.9
		✓	✓	41.7	71.9	42.2	26.2	49.2	59.8
Two-stage detector	Faster R-CNN [4]			38.6	64.0	41.1	20.8	47.4	59.1
		✓		41.2	69.3	42.7	24.1	49.1	59.7
		✓	✓	41.7	70.4	43.0	24.3	50.0	60.6
Anchor-free detector	FCOS [55]			34.7	63.7	34.0	17.8	43.2	51.3
		✓		35.5	64.6	34.8	21.3	42.3	51.1
		✓	✓	37.0	66.5	36.7	21.4	45.0	52.9
	CenterNet [53]			38.8	66.5	39.7	18.8	47.6	63.3
		✓		42.6	71.6	43.8	24.9	50.6	63.2
		✓	✓	41.8	70.3	42.9	23.1	50.2	62.8

TABLE 2. Experimental results on the BDD100k dataset. We conducted comparisons across various object detectors (i.e., SSD, RetinaNet, Faster R-CNN, FCOS, and CenterNet) with and without SR and our RDSP. ✓denotes that the corresponding sub-network is integrated into the full network. Red and blue values indicate the best and the second-best scores in each detector, respectively.

Model		SR	RDSP	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
One-stage detector	SSD [2]			24.7	50.4	20.1	10.5	35.2	53.2
		✓		21.6	42.8	19.1	7.7	33.3	51.8
		✓	✓	29.9	61.1	25.4	15.7	38.6	54.8
	RetinaNet [3]			31.9	64.3	27.8	15.9	42.9	56.0
		✓		30.6	61.3	26.8	13.5	43.3	53.4
		✓	✓	34.3	68.4	30.3	18.0	44.9	57.5
Two-stage detector	Faster R-CNN [4]			32.1	65.0	28.0	17.1	43.1	57.1
		✓		30.8	61.9	26.7	13.7	43.7	53.0
		✓	✓	35.7	71.0	31.1	20.3	45.7	58.0
Anchor-free detector	FCOS [55]			27.2	56.7	22.7	12.7	38.2	49.3
		✓		22.3	48.2	17.6	7.6	33.7	40.2
		✓	✓	29.1	60.6	24.3	14.5	39.1	50.6
	CenterNet [53]			33.1	66.5	28.7	16.6	44.0	59.1
		✓		30.5	61.6	26.3	13.8	42.1	58.5
		✓	✓	29.6	59.5	25.7	12.8	42.1	58.3

the CityScapes dataset in the SSD detector are shown in Figure 8. The first and second rows of the figure show that SR enables the detector to detect small objects that are not detected without SR. Rows 3-6 of the figure shows that, while detection with SR detects false positives in regions where people should not be present in terms of their contexts, our proposed method suppresses these false positives by successfully estimating regions where people should not be and masking them. The quantitative results of various detectors on CityScapes are shown in Table 1. Table 1 shows that the use of SR improves the detection performance of small objects, but sometimes has negative impacts on the performance of medium or large object detections. With our proposed RDSP networks, the detection performances are further improved in most cases. However, the results with CenterNet indicate a performance decrease when utilizing RDSP. The discussion of these results is presented later.

Furthermore, we evaluated the effectiveness of SR and RDSP on the BDD100k dataset. The quantitative results are shown in Table 2. Unlike the results on the CityScapes dataset, the table shows a significant performance drop when using SR without RDSP. This is because of the fact that the BDD100k dataset contains more realistic and challenging images, for example, including stronger blur and reflections on the front windshield. In such images, SR tends to generate more severe artifacts, exacerbating the false-positive issue. Nevertheless, even in these challenging cases, our RDSP mitigates the false-positive issue and dramatically improves overall performance.

2) MODEL COMPLEXITY ANALYSIS

We evaluate the model complexity of our proposed method with the SSD detector. The results are shown in Table 3. From this table, we can see that our proposed RDSP achieves performance improvement with relatively reasonable computational costs and parameter counts compared to not using RDSP. Therefore, RDSP can be considered an efficient method for object detection within the SR-based detection framework. However, when compared to the original SSD, our model has significantly higher runtime and parameter counts. This is because our detection framework has independent detection branches at each

scale. While the parallelization of these branches on high-performance computational resources may mitigate runtime, this complexity of SR-based object detection remains a limitation.

3) COMPARISONS WITH STATE-OF-THE-ART DETECTOR

We compare our proposed framework with several state-of-the-art general object detection methods. These state-of-the-art methods are finetuned on the CityScapes dataset based on the COCO-pretrained models and training configurations provided by respective authors. The results are shown in Table 4. The table reveals that even these SoTA methods exhibit lower performance in some metrics. This highlights the difficulties posed by significant size variations and the practical challenges presented by the CityScapes dataset, captured by real car-mounted cameras. Our proposed method outperforms these SoTA methods, especially in the AP_s metric, indicating the effectiveness of utilizing SR in this kind of challenging dataset.

4) COMPARISONS WITH INTERPOLATION-BASED UPSCALING

The previous scale-specific detectors [5], [6] use bicubic interpolation to upscale images. Furthermore, other methods [3], [53] demonstrated that the detection performance can be improved by merging detection results of upscaled images at several factors by bicubic interpolation. Therefore, we also experiment with bicubic interpolation instead of SR. Additionally, we also present experimental results using a simpler interpolation method, linear interpolation.

The results are shown in Table 5. With RDSP, SR outperforms interpolation-based upscaling. This implies that SR has the potential to outperform interpolation when the false positives are suppressed by our proposals. On the other hand, without RDSP, interpolation-based scaling performs slightly better than SR. In addition, we show examples of upscaled image regions that have false positives by SR in Figure 9. From this figure, we can see that the image upscaled by SR is sharp but has unnatural edges, which are not presented in the interpolation-based method. These edges are a potential cause of false positives. Our proposed RDSP suppresses these false positives, and thus takes advantage of SR for small object detection. On the other hand, with interpolation-based methods, it can be observed that the performance slightly degrades when RDSP is used. This is because RDSP masks certain regions of the image, potentially leading to the loss of useful information for detection, although these regions do not have unnatural artifacts. Furthermore, we can see that there is no significant difference in performance between Bicubic interpolation and Blinear interpolation. Since the proposed method trains the detector with upscaled images, it suggests that the subtle differences arising from the choice of interpolation method do not affect detection accuracy.

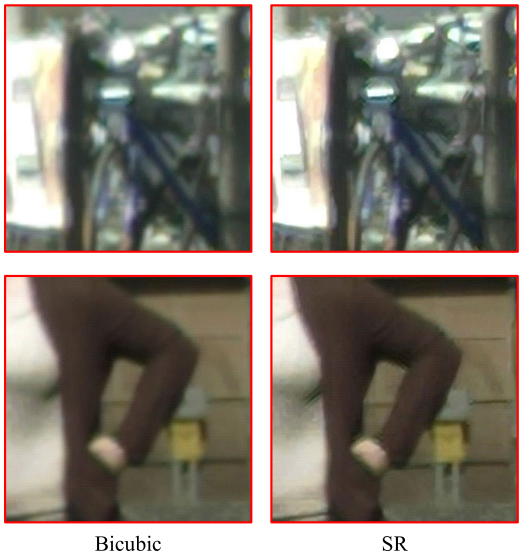

FIGURE 9. The examples of upscaled image regions that have false positives by SR. These images are the results of x4 upscaling using Bicubic and SR. The top images represent a region of the original image size at 48 × 48, while the bottom images represent a region of the original size at 64 × 64. We can see that the SR images are clear but have unnatural edges compared to Bicubic ones.

5) ABLATIONS OF RDSP COMPONENTS

As described in Section III-A, the RDSP consists of the following three components; "Positional embedding," "Scene structure Embedding," and "UNet-like Backbone" (hereinafter called PE, SE, and UB, respectively). We conduct ablation studies to measure the effects of these structural components. For the ablation of UB, we remove the maxpool and up-conv layers from UB since the motivation of UB is to capture the contextual information. This allows comparison with networks that have the same number of layers but cannot capture contexts. The results of ablations are shown in Table 6.

From this table, when only one of the three components is ablated, there is no significant difference in performance compared to using all three components in any case. This suggests that these three components play similar roles in incorporating a global context, although they have different operations. On the other hand, when all three components are ablated, a significant performance drop is observed. This implies that extracting global context is essential for the estimation of RDSP.

6) LEARNING STRATEGIES

In our proposed method, the RDSP is pretrained by our proposed loss function L_{scale}, and then fine-tuned by the combined loss of detection loss L_{det} and proposed loss L_{scale}, as described in Sec III-B. In this section, we present experiments with and without pretraining to validate the effectiveness of pretraining. In addition, we experiment with or without L_{scale} during fine-tuning by changing the weighting parameter for L_{scale} (denoted as λ_{scale}). The

TABLE 3. Comparison of the model complexity of the proposed method.

Model	AP	AP_S	AP_M	AP_L	Runtime[s]	Parameters
SSD	37.3	17.8	46.3	60.0	0.056	36M
SSD + SR	39.5	20.4	48.5	59.2	0.439	153M
SSD + SR + RDSP (ours)	40.3	22.0	49.0	59.8	0.477	158M

TABLE 4. Comparisons with other state-of-the-art object detector. Red and blue values indicate the best and the second-best scores, respectively.

Model	AP	AP_S	AP_M	AP_L
RetinaNet	39.2	22.0	47.8	58.6
RetinaNet + SR	39.5	24.8	48.6	58.9
RetinaNet + SR + RDSP (ours)	41.7	26.2	49.2	59.8
TSP-Det [58]	36.6	17.2	44.4	58.2
Sparse R-CNN [59]	33.7	23.2	45.4	55.5
Anchor DETR [60]	39.1	19.2	45.1	59.8
Deformable DETR [61]	40.1	22.3	48.2	60.1

TABLE 5. Comparisons with interpolation-based upscaling with SSD. ✓denotes that the corresponding architecture is used.

scaling method	RDSP	AP	AP_S	AP_M	AP_L
SR		39.5	20.4	48.5	59.2
	✓	40.3	22.0	49.0	59.5
Bicubic		39.6	20.7	48.3	58.2
	✓	39.6	19.8	48.3	58.3
Bilinear		39.4	20.2	48.6	58.2
	✓	39.7	20.0	48.3	58.0

TABLE 6. Ablation studies for RDSP network with SSD. PE, SE, and UB denote "Positional Embedding," "Scene structure Embedding," and "UNet-like Backbone," respectively. ✓means the corresponding architecture is used. Bold values indicate the performance difference between using all three components and the rest.

PE	CE	UB	AP	AP_S	AP_M	AP_L
✓	✓	✓	40.3	22.0	49.0	59.5
✓	✓		40.0	21.6	49.0	59.2
			−0.3	**−0.4**	**0.0**	**−0.3**
✓		✓	40.0	21.7	49.0	59.6
			−0.3	**−0.3**	**0.0**	**+0.1**
	✓	✓	40.1	21.6	49.1	59.4
			−0.2	**−0.4**	**+0.1**	**−0.1**
			35.3	20.2	45.9	42.1
			−5.0	**−2.0**	**−3.1**	**−17.4**

TABLE 7. Comparison of learning strategies with SSD detector. ✓denotes RDSP network is pretrained.

pretrain	λ_{scale}	AP	AP_S	AP_M	AP_L
✓	1	40.3	22.0	49.0	59.5
	1	39.7	21.1	48.8	58.1
✓	0	39.4	20.5	49.0	58.2
	0	31.2	12.2	48.5	52.8

comparisons of learning strategies are shown in Table 7. From this table, we can see that using L_{scale} during fine-tuning has no positive impact. In contrast, pretraining with L_{scale} improves the detection performance.

TABLE 8. Comparison of detector sharing with SSD.

	AP	AP_S	AP_M	AP_L
Independent detector (proposed)	40.3	22.0	49.0	59.5
Shared detector	36.2	16.8	44.9	59.3

TABLE 9. Comparison of embedding operation.

operation	AP	AP_S	AP_M	AP_L
Add	40.3	22.0	49.0	59.5
Multiply	40.3	21.6	49.0	59.5

7) INDEPENDENT DETECTORS FOR EACH SCALE FACTOR

In our proposed detection pipeline, SR images are fed into independent detectors for each scale factor, while the previous methods [5], [6] utilize shared detectors. This is because SR images upscaled by CNNs have unique features (e.g., checkerboard artifacts) that depend on their upscaling factors, and it is very difficult for a detector to generalize these unique features, such as artifacts. To prevent the detectors from being affected by such unique features, we utilize independent detectors for each scaling factor. To demonstrate the effectiveness of the independent detectors, we experimented with a shared detector on SSD. This detector sharing is done by fixing the detector's weights with the pretrained model described in Section IV-B.

The quantitative results are shown in Table 8. The table shows that detector sharing has a negative effect on the detection performance on all metrics.

8) EMBEDDING OPERATION

In RDSP, positional and scene structure features are embedded using an "Add" operation as described in Sec.III-A. In this section, we explore an alternative embedding method, which is the "Multiply" operation, and provide a comparison between the two approaches. The results of this comparison are shown in Table 9.

From the table, we can see that the "Add" and "Multiply" operations yield very similar performance. This suggests that, fundamentally, the choice between these operations does not significantly impact the embedding's effectiveness.

9) ANALYSIS OF PERFORMANCE DROP IN CENTERNET

As shown in Table 1 and Table 2, our proposed framework performs worse when using CenterNet. We attribute this performance drop to the Deformable Convolution in Centernet. Since Deformable Convolution allows flexible

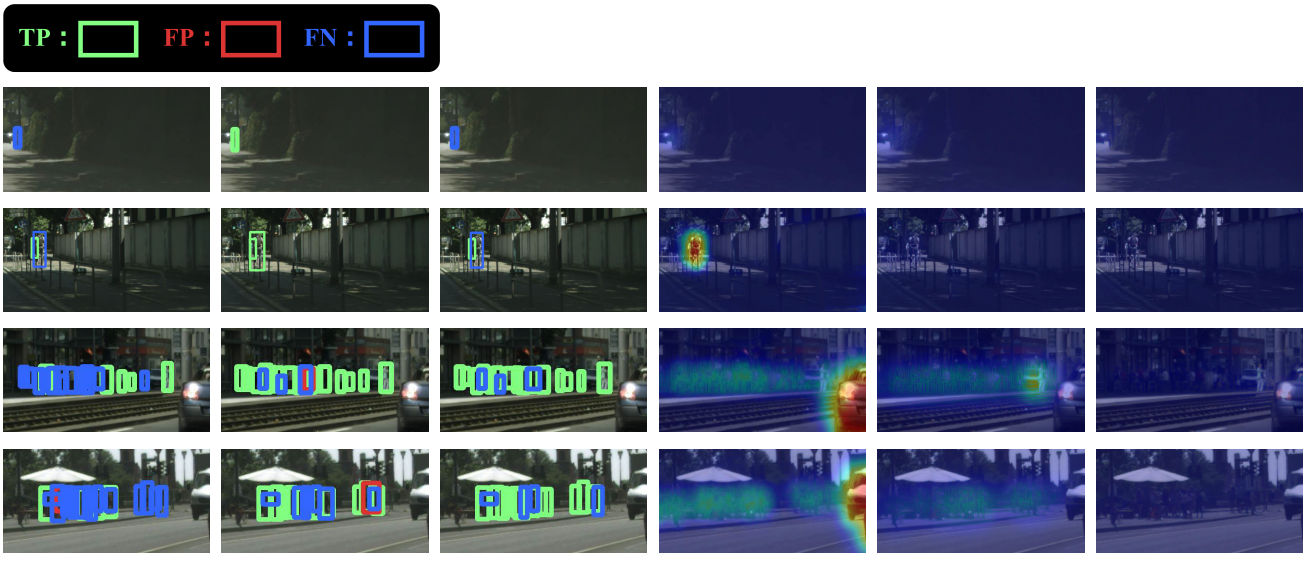

FIGURE 10. The failure case with the proposed RDSP. The predicted bboxes with confidence greater than 50% are visualized. For better visualization, a part of the image is cropped.

TABLE 10. Comparison of CenterNet with and without Deformable Convolution. DC denotes Deformable Convolution, and red values indicate the best scores in each detector.

Model	SR	RDSP	AP	AP_s	AP_m	AP_L
CenterNet			38.8	18.8	47.6	63.3
	✓		42.6	24.9	50.6	63.2
	✓	✓	41.8	23.1	50.2	62.8
CenterNet w/o DC			35.8	17.6	46.5	55.7
	✓		38.6	18.9	47.9	61.0
	✓	✓	39.6	20.1	48.9	61.5

convolution positions and enables the extraction of contextual information, our proposed RDSP imposes a mask on the image, potentially causing the convolution positions of Deformable Convolution to extend into the masked regions. Table 10 shows the comparisons with and without Deformable Convolutions on Centernet. We can see that the performance degradations do not occur on CenterNet without Deformable Convolutions.

10) ANALYSIS OF FAILURE CASE OF PROPOSED METHOD

Figure 10 shows failure cases of the proposed method. The failure cases can be categorized into two types:

- **Limited contextual information.** Since our RDSP relies on contextual information to recognize object scales, heatmap estimation fails when there is limited contextual information. In the examples shown in the 1st and 2nd row of Figure 10, the heatmaps do not activate in regions where it should be estimated as x2 or x4 scales, because the entire scene is dark and challenging to extract contextual information. Such RDSP estimation failures result in undetected objects, which are detected without RDSP.
- **Crowded scenes.** Since our proposed detection framework merges the detection results from each branch using Non-Maximum Suppression (NMS), it struggled to detect overlapping objects. Therefore, while the use of SR and RDSP improves detection accuracy, undetected objects still exist, as shown in the 3rd and 4th row of Figure 10.

Improving the first category could potentially be achieved by providing additional cues beyond the image, such as LiDAR or RADAR data. For the second category, some performance improvement might be achieved through the use of more advanced NMS, such as Soft-NMS [62] or learnable NMS [63].

V. CONCLUDING REMARKS

This paper proposed a method for estimating object-scale proposals for scale-optimized object detection using SR. With images that are rescaled by the appropriate SR scaling factor, an object detector can work better than in the original-size image. A variety of experimental results validated that our proposed RDSP network can capture the rough locations of objects depending on contextual information. We qualitatively and quantitatively verified that object detectors using our scale proposals outperform those without the scale proposals.

Since the proposed method can also be applied to many other computer vision tasks (e.g., human pose estimation, face detection, and human tracking) that capture tiny objects, we would like to extend our proposals to these tasks in future work.

REFERENCES

[1] M. Haris, G. Shakhnarovich, and N. Ukita, "Task-driven super resolution: Object detection in low-resolution images," in *Proc. Int. Conf. Neural Inf. Process.* Sanur, Bali, Indonesia: Springer, 2021, pp. 387–395.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 21–37.
[3] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[4] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[5] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection-SNIP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3578–3587.
[6] B. Singh, M. Najibi, and L. S. Davis, "SNIPER: Efficient multi-scale training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 9333–9343.
[7] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Finding tiny faces in the wild with generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 21–30.
[8] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "SOD-MTGAN: Small object detection via multi-task generative adversarial network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 206–221.
[9] K. Akita, M. Haris, and N. Ukita, "Region-dependent scale proposals for super-resolution in object detection," in *Proc. IEEE 4th Int. Conf. Image Process., Appl. Syst. (IPAS)*, Dec. 2020, pp. 108–113.
[10] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1664–1673.
[11] X. Deng, R. Yang, M. Xu, and P. L. Dragotti, "Wavelet domain style transfer for an effective perception-distortion tradeoff in single image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3076–3085.
[12] A. Lugmayr, M. Danelljan, L. Van Gool, and R. Timofte, "SRFlow: Learning the super-resolution space with normalizing flow," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 715–732.
[13] H. Li, C. Yan, S. Lin, X. Zheng, B. Zhang, F. Yang, and R. Ji, "PAMS: Quantized super-resolution via parameterized max scale," in *Proc. ECCV*, 2020, pp. 564–580.
[14] W. Lee, J. Lee, D. Kim, and B. Ham, "Learning with privileged information for efficient image super-resolution," in *Proc. ECCV*, 2020, pp. 465–482.
[15] X. Luo, Y. Xie, Y. Zhang, Y. Qu, C. Li, and Y. Fu, "LatticeNet: Towards lightweight image super-resolution with lattice block," in *Proc. ECCV*, 2020, pp. 272–289.
[16] Z. Zhang, Z. Wang, Z. Lin, and H. Qi, "Image super-resolution by neural texture transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7974–7983.
[17] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, and J. Zhou, "Structure-preserving super resolution with gradient guidance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7766–7775.
[18] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-GAN," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 284–293.
[19] V. Cornillère, A. Djelouah, W. Yifan, O. Sorkine-Hornung, and C. Schroers, "Blind image super-resolution with spatially variant degradations," *ACM Trans. Graph.*, vol. 38, no. 6, pp. 1–13, Dec. 2019.
[20] J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind super-resolution with iterative kernel correction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1604–1613.
[21] Z. Hui, J. Li, X. Wang, and X. Gao, "Learning the non-differentiable optimization for blind super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2093–2102.
[22] S. Y. Kim, H. Sim, and M. Kim, "Koalanet: Blind super-resolution using kernel-oriented adaptive local adjustment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10606–10615.
[23] Y. Guo, J. Chen, J. Wang, Q. Chen, J. Cao, Z. Deng, Y. Xu, and M. Tan, "Closed-loop matters: Dual regression networks for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5406–5415.
[24] R. Timofte, R. Timofte, S. Gu, J. Wu, and L. Van Gool, "NTIRE 2018 challenge on single image super-resolution: Methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 852–863.
[25] S. Gu et al., "AIM 2019 challenge on image extreme super-resolution: Methods and results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3556–3564.
[26] K. Zhang, S. Gu, and R. Timofte, "NTIRE 2020 challenge on perceptual extreme super-resolution: Methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2045–2057.
[27] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3892–3901.
[28] T. Isobe, X. Jia, S. Gu, S. Li, S. Wang, and Q. Tian, "Video super-resolution with recurrent structure-detail network," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 645–660.
[29] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "TDAN: Temporally-deformable alignment network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3357–3366.
[30] K. C. K. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "BasicVSR: The search for essential components in video super-resolution and beyond," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4945–4954.
[31] K. C. K. Chan, S. Zhou, X. Xu, and C. C. Loy, "BasicVSR: Improving video super-resolution with enhanced propagation and alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5962–5971.
[32] M. Haris, G. Shakhnarovich, and N. Ukita, "Space-time-aware multi-resolution video enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2856–2865.
[33] X. Xiang, Y. Tian, Y. Zhang, Y. Fu, J. P. Allebach, and C. Xu, "Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3367–3376.
[34] G. Xu, J. Xu, Z. Li, L. Wang, X. Sun, and M.-M. Cheng, "Temporal modulation network for controllable space-time video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6384–6393.
[35] Z. Geng, L. Liang, T. Ding, and I. Zharkov, "RSTT: Real-time spatial temporal transformer for space-time video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17420–17430.
[36] S. D. Khan, Y. Ali, B. Zafar, and A. Noorwali, "Robust head detection in complex videos using two-stage deep convolution framework," *IEEE Access*, vol. 8, pp. 98679–98692, 2020.
[37] S. D. Khan and S. Basalamah, "Sparse to dense scale prediction for crowd couting in high density crowds," *Arabian J. Sci. Eng.*, vol. 46, no. 4, pp. 3051–3065, Apr. 2021.
[38] S. D. Khan and S. Basalamah, "Multi-scale person localization with multi-stage deep sequential framework," *Int. J. Comput. Intell. Syst.*, vol. 14, no. 1, p. 1217, 2021.
[39] S. D. Khan and S. Basalamah, "Scale and density invariant head detection deep model for crowd counting in pedestrian crowds," *Vis. Comput.*, vol. 37, no. 8, pp. 2127–2137, Aug. 2021.
[40] A. Mhalla, T. Chateau, H. Maâmatou, S. Gazzah, and N. E. B. Amara, "SMC faster R-CNN: Toward a scene-specialized multi-object detector," *Comput. Vis. Image Understand.*, vol. 164, pp. 3–15, Nov. 2017.
[41] Q. Ye, T. Zhang, W. Ke, Q. Qiu, J. Chen, G. Sapiro, and B. Zhang, "Self-learning scene-specific pedestrian detectors using a progressive latent model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2057–2066.
[42] X. Li, Z. Jie, W. Wang, C. Liu, J. Yang, X. Shen, Z. Lin, Q. Chen, S. Yan, and J. Feng, "FoveaNet: Perspective-aware urban scene parsing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 784–792.
[43] S. Kong and C. Fowlkes, "Recurrent scene parsing with perspective understanding in the loop," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 956–965.
[44] J. Pan and T. Kanade, "Coherent object detection with 3D geometric context from a single image," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2576–2583.
[45] R. Zhang, S. Tang, Y. Zhang, J. Li, and S. Yan, "Scale-adaptive convolutions for scene parsing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2050–2058.
[46] Z. Hao, Y. Liu, H. Qin, J. Yan, X. Li, and X. Hu, "Scale-aware face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1913–1922.
[47] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1522–1530.

[48] Y. Liu, R. Wang, S. Shan, and X. Chen, "Structure inference net: Object detection using scene-level context and instance-level relationships," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6985–6994.
[49] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 213–229.
[50] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "'Convolutional pose machines' is the title of the paper. It should not be modified," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.
[51] W.-C. Hung, Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, and M.-H. Yang, "Scene parsing with global context embedding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2650–2658.
[52] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Munich, Germany: Springer, 2015, pp. 234–241.
[53] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.
[54] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[55] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9626–9635.
[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[57] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.
[58] Z. Sun, S. Cao, Y. Yang, and K. Kitani, "Rethinking transformer-based set prediction for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3591–3600.
[59] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, and P. Luo, "Sparse R-CNN: End-to-end object detection with learnable proposals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14449–14458.
[60] Y. Wang, X. Zhang, T. Yang, and J. Sun, "Anchor DETR: Query design for transformer-based object detection," in *Proc. AAAI*, 2022, pp. 2567–2575.
[61] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. ICLR*, 2021. [Online]. Available: https://iclr.cc/virtual/2021/oral/3448
[62] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS—Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 5562–5570.
[63] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6469–6477.

KAZUTOSHI AKITA was born in Japan, in 1996. He received the B.E. and M.E. degrees from the Toyota Technological Institute, Japan (TTI-J), in 2019 and in 2021, respectively, where he is currently pursuing the Ph.D. degree. His current research interests include object detection, super-resolution, and their coupling.

NORIMICHI UKITA (Member, IEEE) was born in Japan, in 1973. He received the B.E. and M.E. degrees in information engineering from Okayama University, Japan, in 1996 and 1998, respectively, and the Ph.D. degree in informatics from Kyoto University, Japan, in 2001.

From 2001 to 2007, he was an Assistant Professor with the Graduate School of Information Science, Nara Institute of Science and Technology, Japan, where he was an Associate Professor, from 2007 to 2016. In 2016, he was a Professor with the Toyota Technological Institute, Japan.

Prof. Ukita's awards include the Excellent Paper Award of IEICE, in 1999, the Winner Award in the 2018 NTIRE Challenge on Image Super-Resolution, the First Place in the 2018 PIRM Perceptual SR Challenge, the Best Poster Award in the 2019 MVA, and the Best Practical Paper Award in the 2021 MVA.