Task-Driven Super Resolution: Object Detection in Low-resolution Images

Muhammad Haris, Greg Shakhnarovich, *Member, IEEE,* and Norimichi Ukita, *Member, IEEE*

Abstract—We consider how image super-resolution (SR) can contribute to an object detection task in low-resolution images. Intuitively, SR gives a positive impact on the object detection task. While several previous works demonstrated that this intuition is correct, SR and detector are optimized independently in these works. This paper proposes a novel framework to train a deep neural network where the SR sub-network explicitly incorporates a detection loss in its training objective, via a tradeoff with a traditional detection loss. This end-to-end training procedure allows us to train SR preprocessing for any differentiable detector. We demonstrate extensive experiments that show our task-driven SR consistently and significantly improves the accuracy of an object detector on low-resolution images from COCO and PASCAL VOC data set for a variety of conditions and scaling factors.

Index Terms—super-resolution, object detection, end-to-end learning, task network, machine perception, joint optimization

1 Introduction

Image Super-Resolution (SR) belongs to image restoration and enhancement (e.g., denoising and deblurring) algorithms, widely studied in computer vision and graphics. In both communities, the goal is to reconstruct an image from a degenerated version as accurately as possible. The quality of the reconstructed image is evaluated by pixel-based quantitative metrics such as PSNR (peak signal-to-noise ratio) and SSIM (structure similarity) [1]. Recently-proposed perceptual quality [2], [3], [4] can also be employed for evaluation as well as for optimizing the reconstruction model. Relationships between the pixel-based and perceptual quality metrics have been investigated in the literature [5], [6] in order to harmonize these two kinds of metrics. Ultimately, the goal of SR is still to restore an image as well as possible in accordance with criteria in human visual perception.

The connection between SR, and other image restoration tools, and visual recognition is that despite continuing advances in visual recognition, it remains vulnerable to a wide range of image degradation, including low resolution and blur [7], [8]. Image restoration such as SR can serve as an input enhancement step to alleviate this vulnerability. For example, accuracy of many recognition tasks can be improved by deblurring [9], [10], [11], [12] or denoising [13]. SR has been also shown to be effective for such preprocessing for several recognition tasks [14], [15], [16], [17], [18].

Typically, in such applications, the SR is trained in isolation from the downstream task, with the only weak connection through the selection of images to train or fine-tune the SR method (e.g., for character recognition, SR is trained on character images).

We propose to bridge this isolation by explicitly incorporating the objective of the downstream task (such as object detection) into training of an SR module. Figure 1 illustrates the effect of our proposed, task-driven approach to SR. Our proposal (e) generated from a low-resolution (LR) image (b) can successfully bring recognition accuracy close to the score of their original high-resolution (HR) image (a).

Our approach is motivated by two observations:

SR is ill-posed: Many possible HR images when downsampled produce the same LR image. We expect that the additional cue given by the downstream task objective such as detection may help guide the SR solution.

Human perception and machine perception differ: It is known that big differences are observed between human and machine perceptions, in particular, with highly-complex deep networks. This is perhaps best exemplified by adversarial images [19], [20], [21] that can "fool" machine perception but not human. Thus, if our goal is to super-resolve an image in part for machine perception, we believe it is prudent to explicitly "cater" to the machine perception when learning SR.

The two SR images in Fig. 1 (d) and (e) illustrate these points. Both look similar to the human eye, but the detection results differ between these two SR images. Here, two objects (i.e., a person and a motorbike) are detected successfully only in (e). Furthermore, the conventional measure of reconstruction quality (PSNR) fails to capture the difference, assigning higher value to (d) which yields to much worse detection results.

The main contributions of this paper are:

- An approach to SR that uses the power of end-to-end training in deep learning to combine low-level and high-level vision objectives, leading to what we call *Task-Driven Super Resolution* (TDSR). As a means of increasing robustness of object detection to LR inputs, this approach provides results substantially better than other SR methods, and is potentially applicable to a broad range of low-level image processing tools and high-level tasks.
- A novel view of SR, explicitly acknowledging the generative or semantic aspects of SR in high scaling factors, which we hope will encourage additional work in the community to help further reduce the gap between low-level and high-level vision.

- *M. Haris and N. Ukita are with Toyota Technological Institute (TTI), Nagoya, Japan, 468-8511.*
  E-mail: {mharis, ukita}@toyota-ti.ac.jp
- *G. Shakhnarovich is with TTI at Chicago, US. E-mail: greg@ttic.edu*

Manuscript received -; revised -.

(a) HR (b) LR (c) Bicubic SR PSNR: 21.26 (d) SR (no task) PSNR: 22.02 (e) TDSR (proposed) PSNR: 21.54

Fig. 1. Scale sensitivity in object detection and the effectiveness of our proposed method (i.e., end-to-end learning in accordance with the mutual improvement of SR and object detection tasks). Images shown in the top row show (a) an original high resolution image, (b) its low-resolution image (here $1/8$-size, padded with black), (c) SR image obtained by bicubic interpolation, (d) SR image obtained by the SR model optimized with no regard to detection, and (e) SR image obtained by our proposed task-driven SR method, using the same model as in (d). For each of the reconstructed HR images, we also report PSNR w.r.t. the original. Despite ostensibly lower PSNR, the TDSR result recovers the correct detection results with high scores, in this case even suppressing a false detection present in the original HR input.

2 Related Work

While there has been much work on SR and on evaluating and improving some measurement of perceptual quality of images, comparatively little work exist on optimizing image restoration tools for machine perception.

2.1 Image quality assessment

Image restoration and enhancement require appropriate quality assessment metrics both for evaluation and (when machine learning is used) as training objectives. As mentioned in Sec. 1, PSNR and SSIM [1] are widely used as such metrics, focusing on comparing a reconstructed/estimated image with its ground truth image. There exist methods for quality assessment that do not require a reference ground truth image [22], [23], including some that use deep neural networks to learn the metrics [24], [25].

Several quality assessment metrics [26], [27], [28] have been evaluated specifically for SR, including no-reference metrics [29]. However all of these metrics are a proxy for (assumed or approximated) human judgment perceptual quality, and do not consider high-level visual tasks such as recognition.

Some task-dependent quality assessment metrics have been proposed for certain tasks, including biometrics [30], face recognition [31], and object recognition [32], showing improvements vs. the task-agnostic metrics. None of them, however, have been used in a joint learning framework with the underlying image enhancement such as SR.

2.2 Image Super Resolution

A huge variety of image SR techniques have been proposed; see survey papers [33], [34], [35] for more details. While self-contained SR is attractive (e.g., self-similarity based SR [36], [37], [38]), most recent SR algorithms utilize external training images for higher performance; for example, exemplar based [39], [40], [41], regression based [42], [43], and web-retrieval based [44]. The effectiveness of using both self and external images is explored in [45], [46].

Like other vision problems, SR has benefited from recent advances in deep convolutional neural networks (DCNNs). SRCNN [47] enhances the spatial resolution of an input LR image by hand-crafted upsampling filters. The enlarged image is then improved by a DCNN. Further improvements are achieved with more advanced architectures, introducing residual connections [48], [49] and recursive layers [50], however the use of the hand-crafted upsampling filters remains an impediment. That can be alleviated by embedding an upsampling layer into a DCNN [51], [52], [53]. Progressive upsampling [54] is also effective for leveraging information from different scales. By sharing the SR features at different scales by iterative forward and backward projections, DBPN-SR [55] enables the networks to preserve the HR components by learning various up- and down-sampling operators while generating deeper features.

While deep features provided by DCNNs allow us to preserve clear high-frequency photo-realistic textures, it is difficult to completely eliminate blur artifacts. This problem has been addressed by introduction of novel objectives, such as perceptual similarity [2], [3] and adversarial losses [56], [57]. Finally, the two ideas can be combined, incorporating perceptual similarity into generative adversarial networks (GANs) in SRGAN [58].

In contrast to prior work, we explicitly incorporate the objective of a well defined, discriminative task (such as detection) into the SR framework.

2.3 Object detection

Most state-of-the-art object detection algorithms extract or evaluate object proposals (e.g., bounding boxes) [59], [60], [61], [62], [63] within a query image and evaluate the "objectness" of each bounding box for object detection, using DCNN features computed or pooled over each box. In many recent models, the mechanism for producing candidate boxes is incorporated into the network architecture [64].

Unlike approaches using object proposals, SSD [65] and YOLO9000 [66] use pre-set default boxes (a.k.a. anchor boxes) covering a query image. The objectness score is computed for each object category in all boxes while its spatial parameters (e.g., location, scale, and aspect ratio) are optimized. This streamlines the computation at test time and produces extremely fast, as well as accurate, detection framework.

2.4 Detection of small objects

One of the remaining problems in computer vision, such as object detection and scene parsing, is to detect small objects. This issue has been investigated by [67], [68], [69], [70]. Most of these methods proposed context-aware network by re-scaling the input to several resolutions then training the networks at each

resolution or proposing a mechanism to select the pooling field size to preserve the small details. Here we consider an alternative: transform the LR images into HR images using SR. So that, instead of designing more LR friendly detector, we can try to make LR images "look like HR image", for which we have plenty of examples, in the hope that the existing detector "used to HR" will then be able to detect objects. In other words, rather than improve the detector, we pre-process the input to make it more amenable to the detector as is.

Recently, some techniques have been proposed to solve the problems of small object detection using SR [71], [72], [73]. However, the proposed method differs from these techniques in two aspects. First, we show that traditional detection loss can be used to improve machine perception on SR network rather than using additional mechanisms such as adversarial training or two sub-problems optimization which is hard to train. Second, we also show how to explicitly train the model not only for SR but also for joint SR + denoising/deblurring to handle more difficult scenarios where the image are afflicted by additional sources of corruption such as blur and noise.

2.5 Connections to generative models

There is also an interesting connection between our approach and the gradient-based adversarial images [19] as well as the popular "neural art" technique called DeepDream [74]. In both of those, an input image (at full resolution) is modified using gradient descent with the objective to achieve certain output for an image classification network. For adversarial images the goal is to make the network predict an incorrect class, while in DeepDream the goals are aesthetic.

3 Task Driven Super-resolution

Our method relies on two building blocks: an SR network S and a task network D as shown in Fig. 2. The SR network maps an LR image x^l to an HR image x^h producing an SR image $x^{sr} = S(x^l; \theta_{SR})$, where θ_{SR} denotes all the parameters of the SR network. The task network takes an image x and outputs a (possibly structured) prediction $\widehat{y} = D(x; \theta_D)$. We refer to these predictors as "networks" because they are likely to be deep neural networks. However our approach does not presume anything about S and D beyond differentiability for training the whole network with an end-to-end learning scheme.

We assume that the task network D has been trained and its parameters θ_D remain fixed throughout training (and will, for brevity, be omitted from notation).

Our method is applicable to any task network. It can be used for a variety of tasks, for example, depth estimation or semantic segmentation. However, in this paper, we restrict our attention to the object detection task, in which $\widehat{y}$ consists of a set of scored bounding boxes for given object classes.

3.1 Component networks

We use the recently proposed Deep Back-Projection Networks (DBPN) [55] as the SR component. The DBPN achieve state of the art or competitive results on standard SR benchmarks, when trained with the MSE reconstruction loss

$$L_{rec}\left(x^h, x^{sr}\right) = \frac{1}{N}\sum_{i=1}^{N}(x_i^h - x_i^{sr})^2 \tag{1}$$

where i ranges of the N pixel indices in the HR image x^h.

As the detector, we use the Single Shot MultiBox Detector (SSD) [65]. The SSD detector works with a set of default bounding boxes, covering a range of positions, scales and aspect ratios; each box is scored for presence of an object from every class. Given the ground truth for an image x, B is the number of matched default boxes to the ground truth boxes y. These matched boxes form the predicted detections $\widehat{y}(x)$. The task (detection) loss of SSD is combined of confidence loss and localization loss:

$$L_{task}(y, \widehat{y}(x)) = \frac{1}{B}\left[L_{conf}(y, \widehat{y}(x)) + \lambda L_{loc}(y, \widehat{y}(x))\right] \tag{2}$$

The confidence loss L_{conf} penalizes incorrect class predictions for the matched boxes. The localization loss L_{loc} penalizes displacement of boxes vs. the ground truth, using smooth L_1 distance. Both losses in (2) are differentiable with respect to their inputs.

Importantly, every default bounding box in SSD is associated with a set of cells in feature maps (activation layers) computed by a convolutional neural network. As a result, since the loss in (2) decomposes over boxes, it is a differentiable function of the network activations and thus a function of the pixels in the input image, allowing us to incorporate this task loss in the TDSR objective described below.

Both of our chosen component networks have code made publicly available by their authors, and can be trained end to end, providing a convenient testbed for our approach; many other choices are possible, in particular for the detector component, but we do not explore them in this paper.

3.2 Task driven training

Normally, learning-based SR systems are trained using some sort of reconstruction loss L_{rec}, such as mean (over pixels) squared error (MSE) between x^h and x^{sr}. In contrast, the detector is trained with L_{task} intended to improve the measure of its accuracy, typically measured as the average precision (AP) for one class, and the mean AP (mAP) over classes for the entire data set.

Let x be the image with detection ground truth labels y, and let $\downarrow(\cdot)$ denote downscaling of an image by a fixed factor. We propose the compound loss, which on the example (x, y) is given by

$$\begin{aligned} L(x, y; \theta_{SR}) = &\alpha L_{rec}\left(x, S(\downarrow(x); \theta_{SR})\right) + \\ &\beta L_{task}\left(y, D(S(\downarrow(x); \theta_{SR}))\right) \end{aligned} \tag{3}$$

where α and β are weights determining relative strength of the reconstruction loss and the detection loss. Under the assumption that both S and D are differentiable, we can use the chain rule, and compute the gradient of L_{task} with respect to its input, the super-resolved $\downarrow(x)$. Then this per-pixel gradient is combined with the per-pixel gradient of the reconstruction loss L_{rec}. The SR parameters θ_{SR} are then updated using standard back-propagation from this combined gradient:

$$\begin{aligned} &\alpha\frac{\partial}{\partial\theta_{SR}}L_{rec}\left(x, S(\downarrow(x); \theta_{SR})\right) + \\ &\qquad\beta\frac{\partial L_{task}\left(y, D(S(\downarrow(x)))\right)}{\partial S(\downarrow(x))}\frac{\partial S(\downarrow(x))}{\partial\theta_{SR}} \end{aligned} \tag{4}$$

3.3 Interpretation

As mentioned in Section 1, SR is an ill-posed problem. At sufficiently high upscaling factors, it resembles (conditional) image

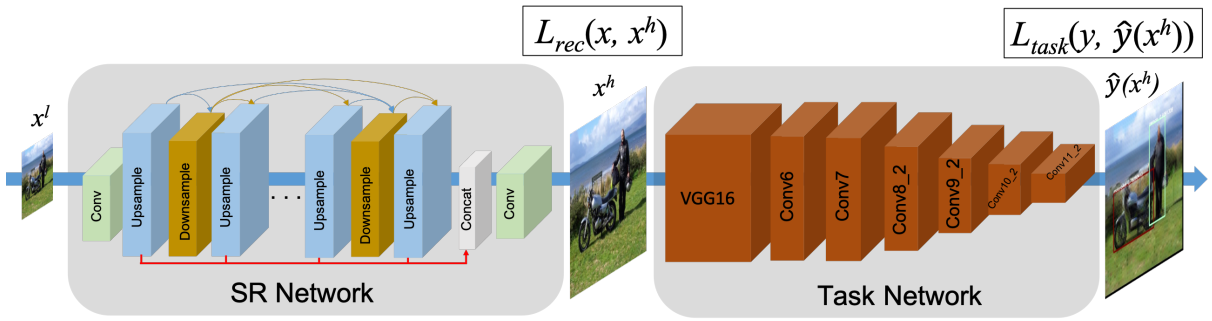


Fig. 2. Network Architecture. DBPN [55] as an SR network and SSD [65] as a task network concatenate to perform end-to-end training.

generation more than image *restoration*, since a large amount of information destroyed in the downscaling process must be "hallucinated". Most current image generation methods, such as GANs or autoencoders, either do not explicitly regard the semantic content of the generated image, or "hardcode" it into the generator by training only on images of a specific class. Our objective (3) encourages the image to both look good to a human (similar to the original) and look correct to the machine (yield the same recognition results). The values of α and β control this tradeoff. With $\alpha \gg \beta$, we effectively ignore the downstream task, and get the traditional, MSE-driven SR learning, with the limitations for downstream detection discussed in Section 2 and demonstrated in Section 4.

With $\beta \gg \alpha$ we effectively ignore the original HR image, and the objective is purely semantic. In this case, intuitively, if the "SR" method were to simply paste a fixed canonical object of the correct class at the appropriate location and scale in the image, and the detector correctly picks up on these objects, we get a perfect value of the task loss. However, in this hypothetical scenario we would in effect replace the SR with a LR detector. That of course would bring up back to the original challenges of LR detection. We also would not get the extra benefit of creating human-interpretable intermediate HR image, connected to the original LR input.

We expect the optimal tradeoff to be somewhere between these scenarios, incorporating meaningful contributions from both the reconstruction and the detection objectives. The precise "mixing" of the two is subject to algorithm design, as detailed in Section 4.2.

4 Experimental Results

4.1 Implementation Details

Base networks DBPN [55] constructs mutually-connected up- and down-sampling layers each of which represents different types of image degradation and HR components. The stack of up- and down- projection units creates an efficient way to iteratively minimize the reconstruction error, to reconstruct a huge variety of SR features, and to enable large scaling factors such as $8\times$ enlargement. We used the setting recommended by the authors: "a 8×8 convolutional layer with four striding and two padding" and "a 12×12 convolutional layer with eight striding and two padding" are used for $4\times$ and $8\times$ SRs, respectively, in order to construct a projection unit. Here, we use D-DBPN which is one of DBPN variants. For object detection, we use SSD300 where the input size is 300×300 pixels. The network uses VGG16 through `conv5_3` layer, then uses `conv4_3`, `conv7` (`fc7`), `conv8_2`, `conv9_2`, `conv10_2`, and `conv11_2` as feature maps to predict the location and confidence score of each detected object. The code for both networks are publicly accessible in the internet.

Datasets We initialized all experiments with DBPN model pre-trained on the DIV2K data set [75], made available by the authors of [55]. We used SSD network pretrained on PASCAL VOC0712 `trainval` and MSCOCO `train2017`. When fine-tuning DBPN in our experiments, with or without task-driven objective, we reused PASCAL VOC0712 `trainval` and MSCOCO `train2017`, with data augmentation. The augmentation consists of photometric distortion, scaling, flipping, random cropping that are recommended to train SSD. Test images on VOC2007 `test` and MSCOCO `val2017` were used for testing in all experiments. The input of DBPN was a LR image that was obtained by bicubic downscaling the original (HR, 300×300) image from the data set with a particular scaling factor (i.e., $1/4$ or $1/8$ in our experiments, corresponding to $4\times$ and $8\times$ SR).

Training setting We used a batch size of 6. The learning rate was initialized to $1e-4$ for all layers and decreased by a factor of 10 after 2×10^5 iterations for training runs consisting of 300,000 iterations. For optimization, we used Adam with momentum set to 0.9. All experiments were conducted using PyTorch 0.3.1 on NVIDIA TITAN X GPUs.

4.2 Training schedules

The definition of loss in (3) depends on the values of α and β, and we can consider a number of settings, both static (fixed weights) and dynamic (weights changing through training).

Fine-tune Generally, we assume that S has been trained for SR for a given factor on images from a domain that could be different from the domain of D. We can simply fine-tune SR on the new domain, without incorporating the task loss: $\alpha = 1$, $\beta = 0$.

Balanced We can start with a phase of fine-tuning the SR on reconstruction only ($\alpha = 1$, $\beta = 0$) and then increase β to a non-zero value, introducing task-driven component. Note that the appropriate relative magnitude of β with respect to α will depend not only on the desired tradeoff between the objectives, but also on the relative scale of the two loss functions.

Task only Alternatively, we can forgo the reconstruction driven phase, and fine-tune S with task loss only, $\alpha = 0$, $\beta = 1$.

Gradual Finally, we can gradually increase β, from zero to a high value, training with each value for a number of iterations. We could expect this schedule to provide a more gentle introduction of the task objective, gradually refining the initially purely reconstruction-driven SR.

4.3 Comparison of Training Schedules

Following the discussion in Sec. 4.2, we investigate different settings and schedules for values of α and β that control the reconstruction-detection tradeoff in (3) trained on PASCAL VOC0712. Table 1 shows PSNR and AP for a number of schedules described on the left in $(n : \alpha : \beta)$ format, indicating training for n iterations with the corresponding values of α (weight on reconstruction loss) and β (weight on detection loss); $+$ indicates continuation of training. The schedules are

(a) SR: Baseline using pretrained SR not fine-tuned on Pascal.
(b) SR-FT: Fine-tuned for $100k$ iterations.
(c) SR-FT+: Fine-tuned for $300k$ iterations.
(d) TDSR-0.1: Balanced schedule in which after $100k$ of reconstruction-only training, we introduce detection loss with the constant weight of $\beta = 0.1$.
(e) TDSR-0.01: Same as previous but the $\beta = 0.01$.
(f) TDSR-DET: $\alpha = 0$ so only detection (AP) loss is used to fine-tune SR for $300k$ iterations.
(g) TDSR-Grad: Gradual increase of β to 1 throughout the $300k$ iterations.

The values in the table provide us with multiple observations. First, it helps to fine-tune SR on the new domain (PASCAL VOC), so SR-FT has much higher PSNR and AP than SR. It helps to fine-tune for longer, hence better results with SR-FT+ (in both PSNR and AP), but we start observing diminishing returns. Switching to variants of TDSR, we see a dramatic increase in AP accuracy. As the relative value of β becomes larger, we get additional improvements, but at the cost of a significant decline in PSNR (and as we see in Fig. 3 and in Section 4.8, in visual quality). However, for a certain regime, namely TDSR-0.01, we see a much higher AP than the no-task values, with only a marginal decline in PSNR. We thus identify this schedule as the best based on our experiments. It is also interesting to see that PSNR can help the network to have better mAP, proven by TDSR-0.01 which has higher mAP than TDSR-DET. Finally, the numbers in the table further illustrate that higher PSNR must not correspond to better detection results.

4.4 Performance on VOC and COCO dataset

Table 2 shows detailed results per class for comparing our TDSR method to other SR approaches trained on VOC0712 `trainval` and evaluated on VOC2007 `test`, including the baseline bicubic SR, and a recently proposed state-of-the-art SR method (SRGAN [58]). Comparison to SRGAN is particularly interesting since it uses a different kind of objective (adversarial/perceptual) which may be assumed to be better suited for task-driven SR. Note that all the other SR models were just pretrained, and not fine-tuned on Pascal. We also compared results obtained directly from LR images (padded with black to fit to the pretrained SSD300 detector). It is shown that SR-FT+ successfully to have highest PSNR. However, TDSR overpowered other methods for all classes and boosted the performance of LR images.

Figure 4 and 5 show graphs where the vertical and horizontal axes denote mAP/PSNR and iterations, respectively, on $4\times$ and $8\times$ in balance setting. It shows that the balance setting successfully increases the accuracy (AP) while maintaining a good quality of images (PSNR).

We see that reduction in resolution has a drastic effect on the AP of the detector, dropping it from 75.8 to 41.7 for $4\times$ and 16.6 for $8\times$ as shown in Table 2. This is presumably due to both the actual loss of information, and the limitations of the detector architecture which may miss small bounding boxes. The performance is not significantly improved by non-task-driven SR methods, which in some cases actually harm it further! However, our proposed TDSR approach obtains significantly better results for both scaling factors, and recovers a significant fraction of the detection accuracy lost in LR.

In accordance with VOC results, the results trained on COCO dataset is also shown the effectiveness of TDSR. Table 3 shows detailed result on COCO `eval2017`. TDSR is successfully to increase the accuracy of LR images roughly by 100% and 500% for $4\times$ and $8\times$, respectively and outperform other methods. Figure 6 and 7 show the accuracy of TDSR and other methods for each class. TDSR consistently has better performance than SR-FT+ for most of the classes especially on $8\times$.

4.5 Performance on different variants of DBPN

In our experiments, we use D-DBPN as the default setting for DBPN. Here, we want to show the trade-off using shallower network from other variants of DBPN. Table 4 shows the comparison of D-DBPN and DBPN-S, which is two variants of DBPN, on performing TDSR. DBPN-S, which has a lower capacity than D-DBPN, gets lower AP of 2.97 and 4.92 than D-DBPN on $4\times$ and $8\times$, respectively. Overall, D-DBPN performs better accuracy than DBPN-S on both scale factors especially $8\times$. However, for real-time application, DBPN-S can be a suitable option which can reduce more than half of DBPN's runtime.

4.6 Fine-tuning SSD

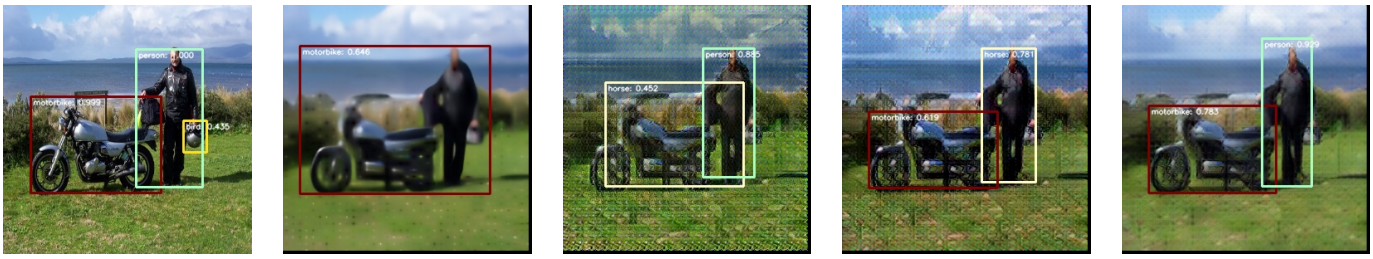
We reiterate that our current focus is on using a fixed pre-trained detector. We also note the common observation that tuning the detector on a modified domain (e.g., degraded images) may impact its performance on the original domain; this is often referred to as forgetting. Also, detecting very small objects is a notoriously difficult task for modern detection frameworks.

However, to show the effectiveness of TDSR even on fine-tuned SSD, we fine-tuned SSD on several SR images, such as Bicubic (SSD-Bicubic), SR-FT+ (SSD-SRFT+), and TDSR (SSD-TDSR) on PASCAL VOC. The results is shown in Table 5. SSD-Pretrained was trained on original HR images. It shows that each domain performs the best on its fine-tuned SSD.

The performance of SSD-Bicubic, which fine-tuned on Bicubic images, is 65.98; unfortunately, the fine-tuned detector now only gets AP of 59.44 on the "original" HR images (down from 75.78). The SSD-SRFT+ achieves 66.58 on SR-FT+ images, but 64.91 on the "original" HR images, down from 75.78. In contrast, with SSD-TDSR, AP on SR images is 67.67 and on original HR images it only drops from 75.78 to 72.99. This suggests: (a) in a specific sense (semantic information accessible to the detector) TDSR is a closer match to HR images; and (b) even if one is willing to have a separate detector for LR images, it is still beneficial to work with TDSR.

4.7 Comparison with Different SR Methods in More Difficult Scenarios

In realistic settings, images are afflicted by additional sources of corruption, which can aggravate the already serious damage from reduction in resolution. In the final set of experiments, we evaluate ours and other methods in such settings. Here, the images (during



(a) HR (b) SR-FT+ PSNR: 22.02 dB (c) TDSR-DET PSNR: 16.63 dB (d) TDSR-Grad PSNR: 19.45 dB (e) TDSR-0.01 PSNR: 21.54 dB

Fig. 3. Comparison on training schedules on $8\times$. PSNR values are for this image only.

TABLE 1
Comparison of training schedules for (3), evaluated on VOC2007 `test`. $n:\alpha:\beta$ indicates training for n iterations with the given α,β values. See text for additional explanations. Red here and in the other tables indicates the best performance.

Setting	HR: 75.78% AP n-iter : α : β	$4\times$ PSNR	$4\times$ AP	$8\times$ PSNR	$8\times$ AP
SR	0k:1:0	22.80	41.9	17.50	10.6
SR-FT	100k:1:0	26.65	52.6	22.77	22.0
SR-FT+	100k:1:0+200k:1:0	26.72	53.6	22.82	22.9
TDSR-0.1	100k:1:0+200k:1:0.1	25.13	61.6	21.08	36.1
TDSR-0.01	100k:1:0+200k:1:0.01	24.06	62.2	22.26	37.5
TDSR-DET	300k:0:1	17.02	61.0	16.72	37.4
TDSR-Grad	100k:1:0+70k:1:0.01+70k:1:0.1+60k:1:1	21.80	61.5	19.78	37.2

TABLE 2
VOC2007 `test` detection results on $4\times$ and $8\times$.

Scale	Method	n-iter : w_{td}	PSNR	AP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv
	HR	-	-	75.8	79.3	85.4	74.1	68.9	46.6	83.7	85.5	86.1	59.1	81.3	77.1	83.5	85.2	82.9	77.6	46.7	73.8	79.9	84.8	73.8
$4\times$	LR	-	-	41.7	48.9	46.8	33.5	31.9	10.7	57.7	48.6	55.9	18.5	31.7	50.1	50.2	61.3	54.2	45.0	18.5	32.8	52.3	52.9	33.4
	Bicubic	-	25.30	41.3	50.9	43.9	37.3	22.0	14.5	53.2	53.9	55.8	18.8	35.6	37.9	52.1	56.9	53.5	49.5	18.7	40.3	51.1	41.8	38.5
	SRGAN	-	23.51	44.6	62.2	45.0	37.0	29.3	15.9	63.0	56.7	44.6	26.5	40.4	46.4	47.9	59.2	52.1	53.1	18.1	40.5	56.9	48.6	47.9
	DBPN	-	22.87	41.9	61.3	41.5	34.4	25.4	16.1	57.7	55.1	43.4	28.9	35.6	44.2	40.7	52.4	47.3	50.0	15.6	32.5	59.1	47.0	50.2
	SR-FT	$100k:0$	26.65	52.6	59.5	61.7	44.3	33.5	26.5	65.6	63.8	61.2	36.2	45.1	55.5	55.7	67.6	64.3	59.4	21.8	45.3	65.8	58.6	60.2
	SR-FT+	$100k:1:0+200k:1:0$	26.72	53.6	59.6	62.9	45.0	34.8	28.3	67.3	64.6	60.7	36.7	45.5	57.5	56.4	68.0	67.0	60.0	22.1	47.9	68.0	59.1	60.7
	TDSR	$100k:1:0+200k:1:0.01$	24.06	62.2	70.6	70.1	55.0	49.4	29.8	71.4	71.1	74.4	41.3	62.6	66.4	69.8	76.1	71.7	67.7	32.8	59.9	71.8	70.9	62.0
$8\times$	LR	-	-	16.6	23.8	17.6	12.2	11.3	9.09	24.6	26.1	23.5	6.27	14.3	13.7	20.1	20.5	23.5	20.6	9.53	10.3	16.2	15.0	12.9
	Bicubic	-	21.85	11.2	13.6	9.80	10.9	1.71	9.09	12.3	18.9	22.7	9.09	7.41	9.91	18.8	10.8	16.9	16.1	2.42	9.09	5.67	2.60	16.1
	SRGAN	-	18.72	13.4	27.2	10.1	12.3	9.96	6.13	15.8	15.6	15.6	9.39	9.89	8.16	18.6	11.7	13.0	20.5	9.44	10.8	17.1	6.59	19.9
	DBPN	-	17.50	10.6	25.0	9.09	10.8	9.54	0.80	16.3	14.7	13.6	3.45	9.09	7.56	12.2	9.09	9.49	13.52	1.96	9.09	16.1	4.55	16.69
	SR-FT	$100k:0$	22.77	22.0	32.0	19.3	18.0	10.7	9.60	34.9	34.6	26.4	13.0	14.5	25.1	27.0	22.2	26.9	31.0	9.46	10.9	26.7	18.1	30.3
	SR-FT+	$100k:1:0+200k:1:0$	22.82	22.9	32.3	24.1	19.7	11.4	9.74	34.8	34.6	27.7	13.3	14.5	24.5	26.7	23.3	28.8	31.9	9.58	11.3	30.1	18.4	30.8
	TDSR	$100k:1:0+200k:1:0.01$	22.26	37.5	49.3	40.9	30.9	25.9	11.4	51.6	47.8	45.0	15.2	31.5	44.1	41.9	50.3	45.6	47.0	14.4	30.6	46.3	40.3	39.6

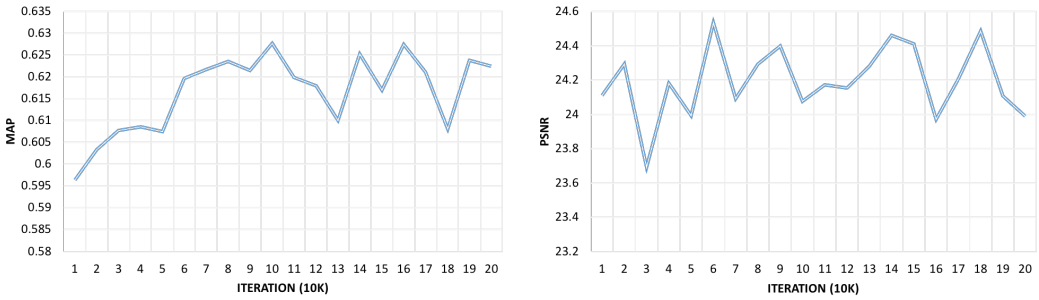


Fig. 4. Convergence $4\times$ which is validated using VOC2007 `test`.

both train and test phases) were also degenerated by blur or noise, prior to downscaling and processing by SR and detector. As with other experiments, we kept the same originally pretrained SSD detector as before.

Blurred Images Every HR image was blurred by Gaussian kernel, $\sigma = 1$. In training the SR network, both in pure SR fine-tuning and in TDSR joint optimization, the objective (L_{rec}) was defined with respect to the original (clean) HR images.

The results of this experiment are shown in Table 6. As with clean images, our proposed method outperforms all other approaches for both scaling factors, even obtaining a small (and likely insignificant) improvement compared to the blurry HR inputs! This application of our method can be thought of as task-driven deblurring by SR.

Noisy Images In a similar vein, we evaluate the SR methods on images affected by Gaussian noise ($\sigma = 0.1$) prior to downscaling. Again, L_{rec} penalizes error w.r.t. the clean HR image.

The AP on noise HR images is 57.3, an almost 20 points drop

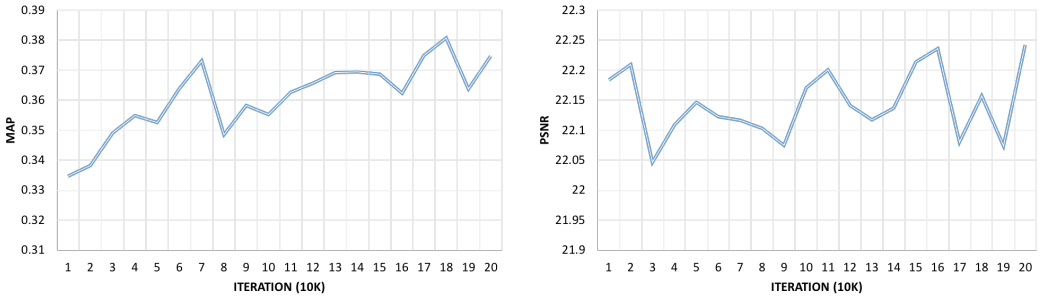


Fig. 5. Convergence $8\times$ which is validated using VOC2007 `test`.

TABLE 3
Results on MSCOCO `val2017`. The bracket values is for ($4\times$: $8\times$) respectively.

	HR	LR	Bicubic	DBPN	SRGAN	SR-FT	SR-FT+	TDSR
AP@[IoU = 0.50 : 0.95 | area= all]	24.2	(8.2 : 1.9)	(8.1 : 2.0)	(1.2 : 0.1)	(0.6 : 0.1)	(13.7 : 4.4)	(14.1 : 4.8)	(16.7 : 9.8)
AP@[IoU = 0.50 | area= all]	42.2	(15.5 : 4.1)	(14.9 : 3.7)	(2.3 : 0.2)	(1.2 : 0.1)	(24.8 : 8.1)	(25.4 : 8.8)	(30.2 : 18.8)
AP@[IoU = 0.75 | area= all]	24.6	(7.9 : 1.7)	(7.8 : 1.9)	(1.1 : 0.0)	(0.6 : 0.0)	(13.7 : 4.3)	(14.0 : 4.7)	(16.7 : 9.2)
AP@[IoU = 0.50 : 0.95 | area= small]	7.2	(0.2 : 0.0)	(0.9 : 0.1)	(0.1 : 0.0)	(0.1 : 0.0)	(2.0 : 0.3)	(2.2 : 0.3)	(2.7 : 0.7)
AP@[IoU = 0.50 : 0.95 | area= medium]	26.7	(3.8 : 0.4)	(6.5 : 1.2)	(0.9 : 0.0)	(0.4 : 0.1)	(12.8 : 3.3)	(13.2 : 3.6)	(15.8 : 6.7)
AP@[IoU = 0.50 : 0.95 | area= large]	39.4	(19.9 : 4.7)	(17.6 : 5.2)	(2.7 : 0.1)	(1.5 : 0.1)	(27.2 : 11.0)	(28.0 : 11.4)	(31.0 : 20.8)

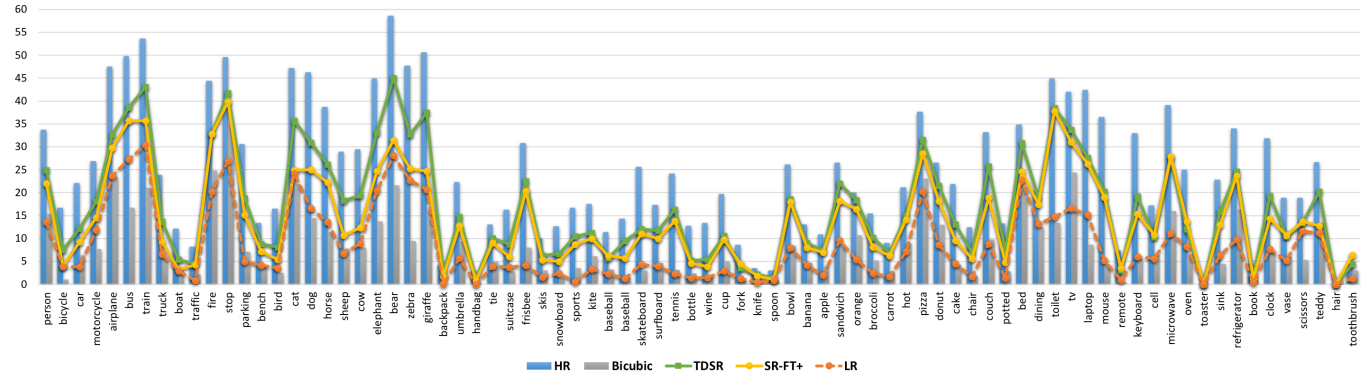


Fig. 6. Result per class from COCO `val2017` on $4\times$, AP at IoU = 0.50 : 0.95.

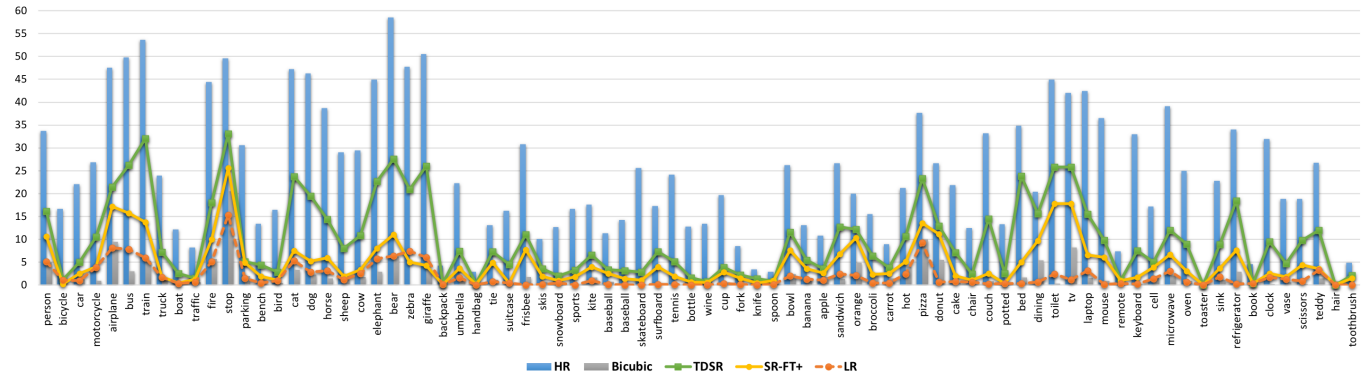


Fig. 7. Result per class from COCO `val2017` on $8\times$, AP at IoU = 0.50 : 0.95.

compared to the clean HR images. The results are shown in Table 7. As with blur, our proposed method outperforms significantly all other approaches for both scaling factors.

4.8 Qualitative Analysis

Figures 8, 9, and 10 show examples of our results compared with those of other methods. The results for SRGAN [58] and SR-FT+ sometimes confuse the detector and recognize it as different object classes, again indicating that optimizing L_{rec} and high PSNR do not necessarily correlate with the accuracy. Meanwhile,

TABLE 4
Average Precision on the variants of DBPN.

	4×		8×	
Input	D-DBPN	DBPN-S	D-DBPN	DBPN-S
SR-FT+	53.60	49.77	22.89	18.10
TDSR	62.25	59.28	37.49	32.57
# parameters (k)	10426	2386	23205	5336
runtime (sec)	0.122	0.054	0.114	0.051

TABLE 5
Average Precision for fine-tuned SSD on different image domain ($4\times$).

Input	SSD-Pretrained	SSD-Bicubic	SSD-SRFT+	SSD-TDSR
HR	75.78	59.44	64.91	72.99
Bicubic	41.32	65.98	48.33	46.67
SR-FT+	53.60	56.45	66.58	58.11
TDSR	62.25	52.54	49.50	67.67

unique pattern that produced by our proposed optimization helps the detector to recognize the objects better. Note that the TDSR does produce, in many images, artifacts somewhat reminiscent of those in DeepDream [74], but those are mild, and are offset by a drastically increased detection accuracy.

5 Conclusions

We have proposed a novel objective for training SR: a compound loss that caters to the downstream semantic task, and not just to the pixel-wise image reconstruction task as traditionally done. Our results, which consistently exceed alternative SR methods in all conditions, indicate that modern end-to-end training enables joint optimization of tasks what has traditionally been separated into low-level vision (super-resolution) and high-level vision (object detection). These results also suggest some avenues for future work. The first is to investigate task-driven SR methods for additional visual tasks, such as semantic segmentation, image captioning, etc. A complementary direction is to extend the task-driven formulation to other image reconstruction and enhancement tools. For instance, we have demonstrated some success in "deblurring by SR", and one can expect further improvement when using a properly designed deblurring network combined with task-driven objectives. Finally, the community may be well served by a continuing quest for better image quality metrics, to replace or augment simplistic reconstruction losses such as mean square error.

TABLE 6
Analysis on blur images of VOC2007 `test`. Note: Original images (HR + Blur) obtained 63.3% AP

Method	n-iter : α : β	4×	8×
LR	-	40.1	16.2
Bicubic	-	42.9	11.8
SR-FT	-	54.7	23.9
SR-FT+	$100k:1:0$+$200k:1:0$	55.5	25.1
TDSR	$100k:1:0$+$200k:1:0.1$	63.8	39.1

TABLE 7
Analysis on noise images of VOC2007 `test`. Note: Original images (HR + Noise) obtained 57.3% AP

Method	n-iter : α : β	4×	8×
LR	-	39.0	14.5
Bicubic	-	21.2	2.84
SR-FT	-	41.5	11.6
SR-FT+	$100k:1:0$+$200k:1:0$	42.7	12.6
TDSR	$100k:1:0$+$200k:1:0.1$	50.1	22.7

References

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[2] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[3] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 658–666.

[4] M. S. Sajjadi, B. Schölkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4501–4510.

[5] P. Hanhart, P. Korshunov, and T. Ebrahimi, "Benchmarking of quality metrics on ultra-high definition video sequences," in *Digital Signal Processing (DSP), 2013 18th International Conference on*. IEEE, 2013, pp. 1–8.

[6] D. Kundu and B. L. Evans, "Full-reference visual quality assessment for synthetic images: A subjective study," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2374–2378.

[7] I. Vasiljevic, A. Chakrabarti, and G. Shakhnarovich, "Examining the impact of blur on recognition by convolutional networks," *arXiv preprint arXiv:1611.05760*, 2016.

[8] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*, 2016.

[9] M. Nishiyama, A. Hadid, H. Takeshima, J. Shotton, T. Kozakaya, and O. Yamaguchi, "Facial deblur inference using subspace analysis for recognition of blurred faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 838–845, 2011.

[10] X. Chen, X. He, J. Yang, and Q. Wu, "An effective document image deblurring algorithm," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 369–376.

[11] M. Hradiš, J. Kotera, P. Zemcík, and F. Šroubek, "Convolutional neural networks for direct text deblurring," in *Proceedings of BMVC*, vol. 10, 2015.

[12] L. Xiao, J. Wang, W. Heidrich, and M. Hirsch, "Learning high-order filters for efficient blind deconvolution of document photographs," in *European Conference on Computer Vision*. Springer, 2016, pp. 734–749.

[13] S. Milani, R. Bernardini, and R. Rinaldo, "Adaptive denoising filtering for object detection applications," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*. IEEE, 2012, pp. 1013–1016.

[14] D. Dai, Y. Wang, Y. Chen, and L. Van Gool, "Is image super-resolution helpful for other vision tasks?" in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–9.

[15] P. H. Hennings-Yeomans, S. Baker, and B. V. Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[16] P. H. Hennings-Yeomans, B. V. Kumar, and S. Baker, "Robust low-resolution face identification and verification using high-resolution features," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 33–36.

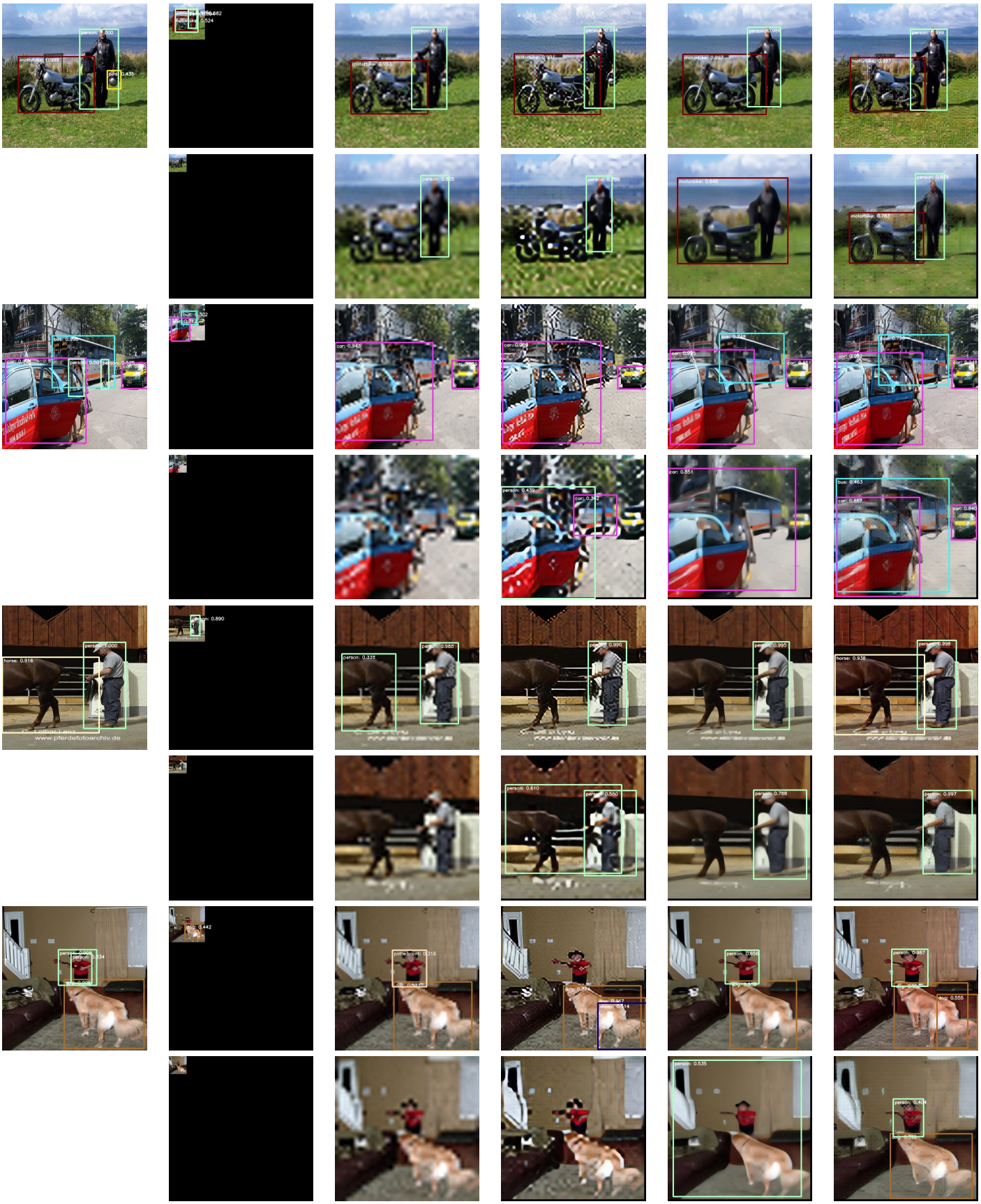
[17] S. Shekhar, V. M. Patel, and R. Chellappa, "Synthesis-based recognition of low resolution faces," in *Biometrics (IJCB), 2011 International Joint Conference on*. IEEE, 2011, pp. 1–6.

[18] E. Bilgazyev, B. A. Efraty, S. K. Shah, and I. A. Kakadiaris, "Sparse representation-based super resolution for face recognition at a distance." in *BMVC*, 2011, pp. 1–11.

[19] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

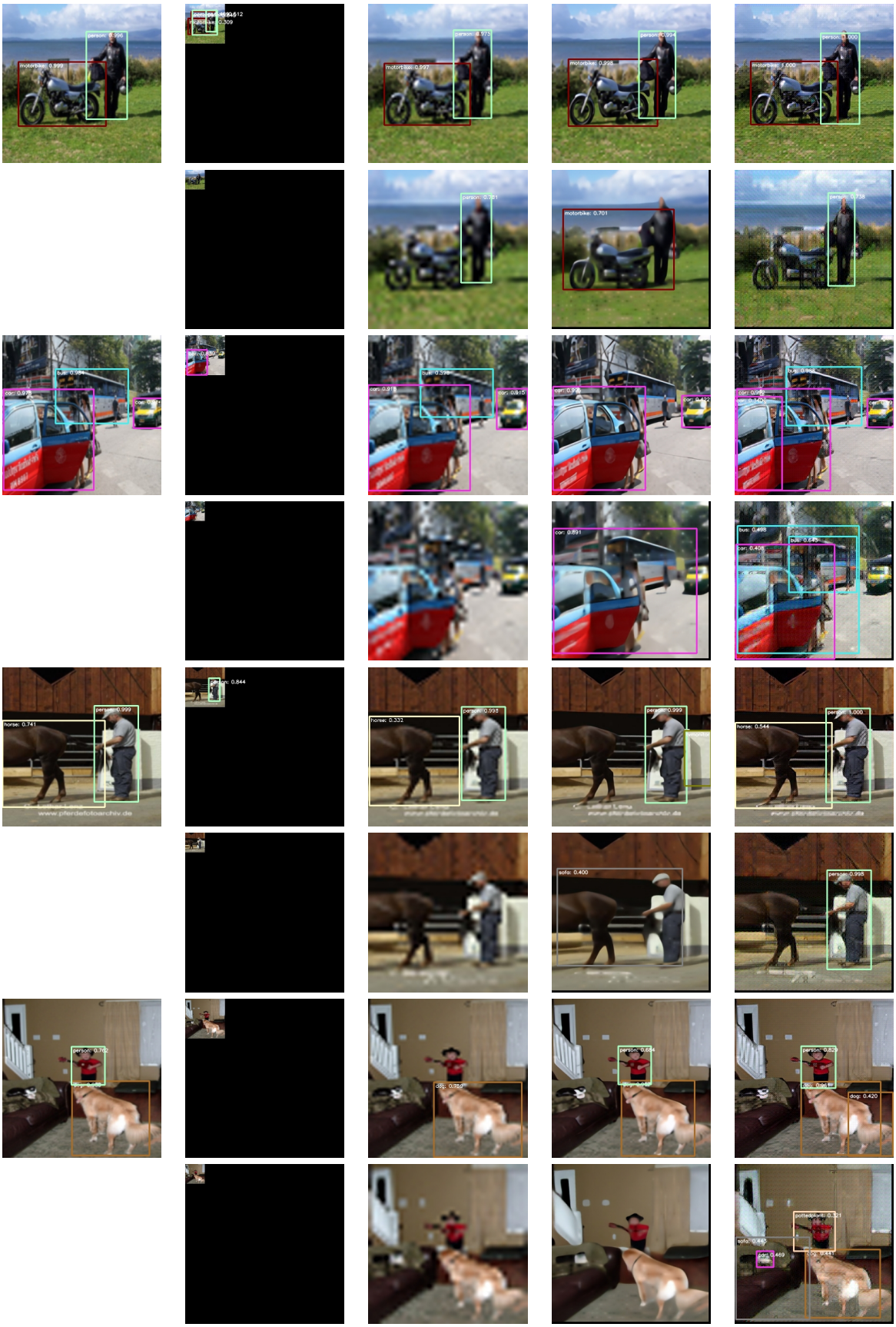
[20] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 427–436.

[21] S. M. Moosavi Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of*



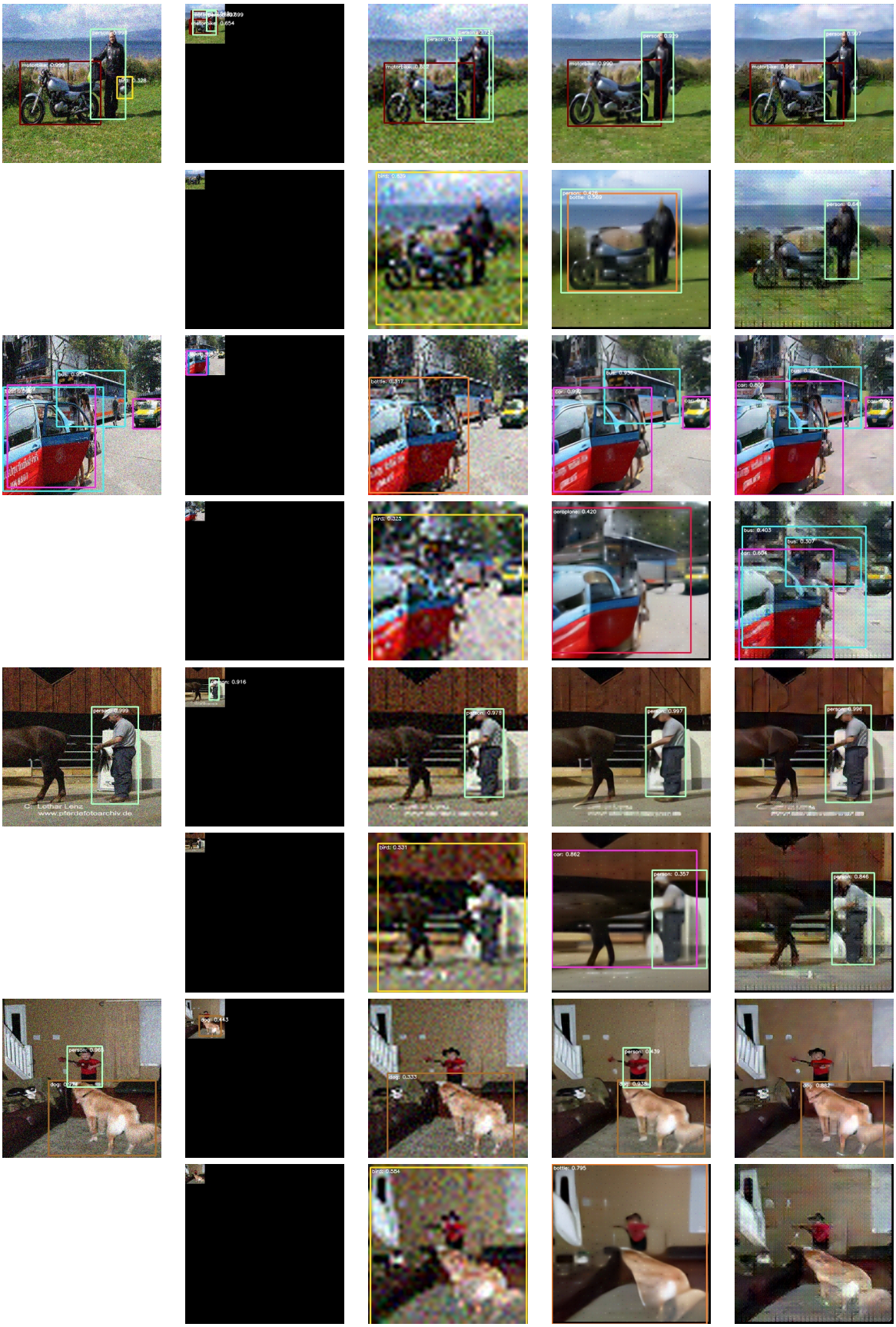
(a) HR (b) LR (c) Bicubic (d) SRGAN [58] (e) SR-FT+ (f) TDSR

Fig. 8. Sample results for $4\times$ (upper row) and $8\times$ (lower row). Zoom in to see detection labels and scores.



(a) HR+Blur (b) LR (c) Bicubic (d) SR-FT+ (e) TDSR

Fig. 9. Sample results on blur images for $4\times$ (upper row) and $8\times$ (lower row). Zoom in to see detection labels and scores.



(a) HR+Noise (b) LR (c) Bicubic (d) SR-FT+ (e) TDSR

Fig. 10. Sample results on noise images for $4\times$ (upper row) and $8\times$ (lower row). Zoom in to see detection labels and scores.

2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[22] H. Luo, "A training-based no-reference image quality assessment algorithm," in *Image Processing, 2004. ICIP'04. 2004 International Conference on*, vol. 5. IEEE, 2004, pp. 2973–2976.

[23] H. Tang, N. Joshi, and A. Kapoor, "Blind image quality assessment using semi-supervised rectifier networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2877–2884.

[24] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.

[25] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2018.

[26] A. R. Reibman, R. M. Bell, and S. Gray, "Quality assessment for super-resolution image enhancement," in *Image Processing, 2006 IEEE International Conference on*. IEEE, 2006, pp. 2017–2020.

[27] H. Yeganeh, M. Rostami, and Z. Wang, "Objective quality assessment for image super-resolution: A natural scene statistics approach," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*. IEEE, 2012, pp. 1481–1484.

[28] Y. Fang, J. Liu, Y. Zhang, W. Lin, and Z. Guo, "Quality assessment for image super-resolution based on energy change and texture variation," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2057–2061.

[29] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Computer Vision and Image Understanding*, vol. 158, pp. 1–16, 2017.

[30] J. Galbally, S. Marcel, and J. Fiérrez, "Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition," *IEEE Trans. Image Processing*, vol. 23, no. 2, pp. 710–724, 2014. [Online]. Available: https://doi.org/10.1109/TIP.2013.2292332

[31] C. G. R. Pulecio, H. D. Benítez-Restrepo, and A. C. Bovik, "Image quality assessment to enhance infrared face recognition," in *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 805–809.

[32] T. Yuan, X. Zheng, X. Hu, W. Zhou, and W. Wang, "A method for the evaluation of image quality according to the recognition effectiveness of objects in the optical remote sensing image using machine learning algorithm," *PLoS One*, vol. 9, no. 1, 2014.

[33] J. D. van Ouwerkerk, "Image super-resolution survey," *Image Vision Comput.*, vol. 24, no. 10, pp. 1039–1052, 2006. [Online]. Available: https://doi.org/10.1016/j.imavis.2006.02.026

[34] K. Nasrollahi and T. B. Moeslund, "Super-resolution: a comprehensive survey," *Mach. Vis. Appl.*, vol. 25, no. 6, pp. 1423–1468, 2014. [Online]. Available: https://doi.org/10.1007/s00138-014-0623-4

[35] C.-Y. Yang, C. Ma, and M.-H. Yang, "Single-image super-resolution: A benchmark," in *European Conference on Computer Vision*. Springer, 2014, pp. 372–386.

[36] C.-Y. Yang, J.-B. Huang, and M.-H. Yang, "Exploiting self-similarities for single frame super-resolution," in *Asian conference on computer vision*. Springer, 2010, pp. 497–510.

[37] T. Michaeli and M. Irani, "Nonparametric blind super-resolution," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 945–952.

[38] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5197–5206.

[39] R. Timofte, R. Rothe, and L. Van Gool, "Seven ways to improve example-based single image super resolution," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2016, pp. 1865–1873.

[40] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE transactions on image processing*, vol. 21, no. 8, pp. 3467–3478, 2012.

[41] A. Bansal, Y. Sheikh, and D. Ramanan, "Pixelnn: Example-based image synthesis," in *ICLR*, 2018.

[42] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 6, pp. 1127–1133, 2010.

[43] E. Pérez-Pellitero, J. Salvador, J. Ruiz-Hidalgo, and B. Rosenhahn, "Psyco: Manifold span reduction for super resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1837–1845.

[44] H. Yue, X. Sun, J. Yang, and F. Wu, "Landmark image super-resolution by retrieving web images," *IEEE Trans. Image Processing*, vol. 22, no. 12, pp. 4865–4878, 2013. [Online]. Available: https://doi.org/10.1109/TIP.2013.2279315

[45] J. Yang, Z. Lin, and S. Cohen, "Fast image super-resolution based on in-place example regression," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 1059–1066.

[46] Z. Wang, Y. Yang, Z. Wang, S. Chang, J. Yang, and T. S. Huang, "Learning super-resolution jointly from external and internal examples," *IEEE Trans. Image Processing*, vol. 24, no. 11, pp. 4359–4371, 2015. [Online]. Available: https://doi.org/10.1109/TIP.2015.2462113

[47] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[48] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp. 1646–1654.

[49] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[50] J. Kim, J. Kwon Lee, and K. Mu Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1637–1645.

[51] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *European Conference on Computer Vision*. Springer, 2016, pp. 391–407.

[52] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.

[53] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[54] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[55] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[56] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[57] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 318–333.

[58] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[59] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013. [Online]. Available: https://doi.org/10.1007/s11263-013-0620-5

[60] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision*. Springer, 2014, pp. 391–405.

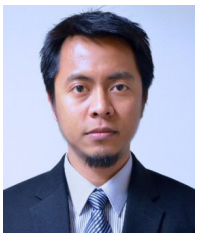
[61] R. B. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015, pp. 1440–1448. [Online]. Available: https://doi.org/10.1109/ICCV.2015.169

[62] J. H. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 814–830, 2016. [Online]. Available: https://doi.org/10.1109/TPAMI.2015.2465908

[63] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, 2016. [Online]. Available: https://doi.org/10.1109/TPAMI.2015.2437384

[64] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[65] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.
[66] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," *arXiv preprint*, vol. 1612, 2016.
[67] P. Hu and D. Ramanan, "Finding tiny faces," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 1522–1530.
[68] 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 *Proceedings of 2017 IEEE Conference on Computer Vision (ICCV)*, 2017.
[69] S. Kong and C. Fowlkes, "Recurrent scene parsing with perspective understanding in the loop," in *Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
[70] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li, "Robust multi-resolution pedestrian detection in traffic scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3033–3040.
[71] Y. Zhang, Y. Bai, M. Ding, and B. Ghanem, "Sod-mtgan: Small object detection via multi-task generative adversarial network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 206–221.
[72] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Finding tiny faces in the wild with generative adversarial network," *CVPR. IEEE*, 2018.
[73] X. Zhao, W. Li, Y. Zhang, and Z. Feng, "Residual super-resolution single shot network for low-resolution object detection," *IEEE Access*, 2018.
[74] A. Mordvintsev, M. Tyka, and C. Olah, "Inceptionism: Going deeper into neural networks," https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html, June 2015.
[75] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.



Muhammad Haris Muhammad Haris received S. Kom (Bachelor of Computer Science) from the Faculty of Computer Science, University of Indonesia, Depok, Indonesia, in 2009. Then, he received the M. Eng and Dr. Eng degree from Department of Intelligent Interaction Technologies, University of Tsukuba, Japan, in 2014 and 2017, respectively, under the supervision of Dr. Hajime Nobuhara. Currently, he is working as postdoctoral fellow in Intelligent Information Media Laboratory, Toyota Technological Institute with Prof. Norimichi Ukita. His main research interests are low-level vision and image/video processing.

Greg Shakhnarovich Greg Shakhnarovich has been faculty member at TTI-Chicago since 2008. He received his BSc degree in Computer Science and Mathematics from the Hebrew University in Jerusalem, Israel, in 1994, and a MSc degree in Computer Science from the Technion, Israel, in 2000. Prior to joining TTIC Greg was a Postdoctoral Research Associate at Brown University, collaborating with researchers at the Computer Science Department and the Brain Sciences program there. Greg's research interests lie broadly in computer vision and machine learning.

Norimichi Ukita Norimichi Ukita is a professor at the graduate school of engineering, Toyota Technological Institute, Japan (TTI-J). 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. After working for five years as an assistant professor at NAIST, he became an associate professor in 2007 and moved to TTIJ in 2016. He was a research scientist of Precursory Research for Embryonic Science and Technology, Japan Science and Technology Agency (JST), during 2002 - 2006. He was a visiting research scientist at Carnegie Mellon University during 2007-2009. He currently works also at the Cybermedia center of Osaka University as a guest professor. His main research interests are object detection/tracking and human pose/shape estimation. He is a member of the IEEE.