RestoreDet: Degradation Equivariant Representation for Object Detection in Low Resolution Images

Ziteng Cui¹, Yingying Zhu², Lin Gu³,⁴, Guo-Jun Qi⁵, Xiaoxiao Li⁶, Peng Gao⁷, Zenghui Zhang¹, Tatsuya Harada⁴,³
¹Shanghai Jiao Tong University, ²University of Texas at Arlington, ³RIKEN AIP, ⁴The University of Tokyo ⁵Innopeak Technology, ⁶The University of British Columbia ⁷Shanghai AI Laboratory

Abstract

Image restoration algorithms such as super resolution (SR) are indispensable pre-processing modules for object detection in degraded images. However, most of these algorithms assume the degradation is fixed and known a priori. When the real degradation is unknown or differs from assumption, both the pre-processing module and the consequent high-level task such as object detection would fail. Here, we propose a novel framework, RestoreDet, to detect objects in degraded low resolution images. RestoreDet utilizes the downsampling degradation as a kind of transformation for self-supervised signals to explore the equivariant representation against various resolutions and other degradation conditions. Specifically, we learn this intrinsic visual structure by encoding and decoding the degradation transformation from a pair of original and randomly degraded images. The framework could further take the advantage of advanced SR architectures with an arbitrary resolution restoring decoder to reconstruct the original correspondence from the degraded input image. Both the representation learning and object detection are optimized jointly in an end-to-end training fashion. RestoreDet is a generic framework that could be implemented on any mainstream object detection architectures. The extensive experiment shows that our framework based on CenterNet has achieved superior performance compared with existing methods when facing variant degradation situations. Our code would be released soon.

1. Introduction

High level vision tasks (i.e., image classification, object detection, and semantic segmentation) have witnessed great success thanks to the large scale dataset [14, 39, 18]. Images in these datasets are mainly captured by commercial cameras with higher resolution and signal-to-noise ratio (SNR). Trained and optimized on these high-quality images, high-level vision would suffer a performance drop on low resolution [13, 25, 55] or low quality images [42, 54, 49, 1, 57, 12].

To improve the performance of vision algorithms on degraded low resolution images, Dai et al. [13] presented the first comprehensive study advocating pre-processing images with super resolution (SR) algorithms. Other high-level tasks like face recognition [74], face detection [3], image classification [59, 42] and semantic segmentation [55], also benefit from the restoration module to extract more discriminate features.

Most existing enhancement methods, especially SR algorithms [4, 66, 65], assume target images are from a known and fixed degradation model [17, 40]:

\[ t(x) = (x \ast k) \downarrow_s + n, \]

where \( t(x) \) and \( x \) denote the degraded low resolution (LR) image and original high resolution (HR) input respectively. \( k \) is the blur kernel while \( \downarrow_s \) is the down-sampling operation with ratio \( s \). \( n \) is the additive noise. However, the performance of these enhancement algorithms would decline severely when the real degradation deviates from the assumption [23]. To make it worse, since these restoration methods are human visual perception oriented, they are often inadequate for machine perception tasks such as object detection [12, 62, 52].

Instead of explicitly enhancing an input image with a restoration module under strict assumptions, we exploit the intrinsic equivariant representation against various resolutions and degradation status. Based on the encoded representation shown in Fig.1, we propose RestoreDet, an end-to-end model for object detection in degraded LR images. To capture the complex patterns of visual structures, we utilize groups of downsampling degradation transformations as the self-supervised signal [17, 68]. During the training, we generate a degraded LR image \( t(x) \) from the original HR image \( x \) through a random degradation transformation \( t \). As shown in Fig.1, this pair of images are fed into the encoder

*Corresponding author.
We propose a novel framework, RestoreDet, to detect objects in degraded low resolution images. To achieve this, we learn the degradation equivariant representation that captures the dynamics of feature representations under diverse resolutions and degradation types. RestoreDet is a generic framework that could be integrated into standard object detection architecture.

- Our method takes the strength of advanced super resolution (SR) research by training an arbitrary resolution restoration decoder (ARRD) that reconstructs the high resolution details. Furthermore, by optimizing the representation learning and detection in a unified end-to-end training framework, the representation preserves the intrinsic visual structure that is discriminative for detection.

- We evaluate our method on two mainstream public datasets KITTI [22] and MS COCO [39]. The experiment results also show that our method has achieved SOTA performance on all low resolution object detection tasks (down 1 \sim 4, down 2, down 4). We will release the source code and results on Github upon the acceptance of this submission.

2. Related Works

2.1. Single Image Super Resolution

The very first CCN-based SR was proposed in Dong et al. [15] with a three-layer neural network. Then Kim et al. [34] extended the depth of network to 20 layers with gradient clipping and residual learning. Batch normalization is later identified to impose the negative effect on the HR reconstruction. By removing this layer, Lim et al. proposed enhanced deep super resolution network (EDSR) [38] that achieves the SOTA in 2017. After ESDR, better SR architectures are designed by integrating the successful deep learning techniques such as Laplacian pyramid structure [35], dense connection [53], residual dense connection [70], graph neural network [72] and so on. Besides designing sophisticated architecture, losses like perceptual loss [31], adversarial loss [36] and targeted perceptual loss [46] are also demonstrated to improve the SR reconstruction quality.

SR algorithms heavily rely on the assumption of degradation model. Much efforts are spent to relax the constraint. Not limited to a fixed integral upsampling scale, Hu et al. [28] proposed a meta-upscaling factor to super-resolve the image with arbitrary scale factor. Recently, Fu et al. [19] proposed a solution based on residual attention network.

To deploy SR for real scenarios, blind SISR methods assume the degradation information is not known. One direction is to convert the problem into non-blind SR by initially estimate the degradation parameters [4]. However, the applied non-blind SR algorithm [2] is very sensitive to the...
error of the degradation estimation. Gu et al. [23] then proposed to iteratively correct the estimated degradation with an iterative kernel correction (IKC) method. Without explicitly estimating degradation parameters, Wang et al. [56] introduced a contrastive loss to design the Degradation-Aware SR network based on the learned representations. Recently, Zhang et al. [65] solved the general blind SISR by designing a practical model considering complex degradations. This model has been demonstrated to cover the degradation space of real images. Therefore, we also adopt this practical model to synthesize various degraded LR images as the self-supervised signal to train our model.

2.2. Image Restoration for Machine Perception

There is sufficient evidence that pre-processing module is an effective in improving the performance of high-level tasks on degraded images [62, 52]. Dai et al. [13] made the first extensive analysis on improving several vision tasks with SR as pre-process. Wang et al. [59] analyzed the effectiveness of SR in image classification task while DSRL [55] improved the semantic segmentation with an additional SR block. Shermeyer and Etten [50] evaluate the effectiveness of a SR pre-process step on aerial image object detection. Recently, Haris et al. [25] jointly optimized the SSD detection loss [43] along with SR sub-network [24] to improve detection performance.

Similarly, the denoising algorithms have also be explored in [42, 32, 49]. Kamann et al. [32] studied the impact of noise and blur on different semantic segmentation methods. Liu et al. [42] combined a denoise network in classifier to improve classification’s performance. Very recently, Mohamed and Gabriel [49] analyzed motion blur to improve detection performance on motion blurry images.

However, most of these existing works assume the degradation parameters such as the downsampling ratio is known. Based on the degradation equivariant representation, our proposed framework is robust to various unknown degradation in real-world scenarios. Without an explicit restoration module, we directly perform the detection on low-dimension encoded features that saves much computational burden.

3. Downsampling Degradation Transformations

In real scenarios, the image may be captured and processed in various ways. To cover these generally unknown operations, it is necessary to select a practical degradation model for the degradation transformation. This model would transform the high resolution (HR) image \( x \) to the low resolution counterpart \( t(x) \) with Eq. 1.

Early enhancement methods assume a simple degradation model where LR is directly downsampled from the HR images without or with simple noise. Instead of dealing with synthetic images, more and more methods now focus on more realistic degradation models. For example, [5, 61] directly train their model on the LR and HR images pair captured by the real camera system. USRNet [66] effectively handled the degradation models with different parameters such as scale factors by unfolding the model-based energy function. Therefore, we adopt the practical degradation model [17, 40, 65] that accounts for diverse degradation in real images.

Our degradation process in Eq.1 mainly comprises of three parts: blur kernel \( k \), down-sampling ratio \( s \) and additional noise \( n \). During training, \( k, s, n \) is randomly sampled from the corresponding distributions as follows:

**Convolution Operation:** Gaussian blur kernel is the most common kernel to blur the image [16, 41, 4, 63]. Here, we choose two Gaussian degradation kernels: isotropic Gaussian kernels \( k_{iso} \) and anisotropic Gaussian kernels \( k_{aniso} \). We also consider None degradation kernel \( k_{none} \). Following [65], the kernel size is uniformly sampled from \( \{7 \times 7, 9 \times 9, ..., 21 \times 21 \} \) and the \( k_{iso}'s \) width is uniformly chosen from \([0, 1, 2, 4]\). For \( k_{aniso} \), the kernel angle is uniformly chosen from \([0, \pi]\) and the longer kernel width is uniformly chosen from \([0.5, 6]\). Here, the transformation decoder \( D_t \) is trained to learn the kernel size. During training, we sample different kernel size, width, and types with the equal probability.

**Down-sampling:** For the down-sampling process, the sampling ratio \( s \) is randomly chosen from uniform distribution \( U(1, 4) \) (e.g. 2.2) while the down-sampling methods is randomly chosen from nearest method \( d^n_{\text{nearest}} \), bilinear method \( d^n_{\text{bilinear}} \), and bicubic method \( d^n_{\text{bicubic}} \). In the actual training phase, the down-sampled resolution must be multiple of the network’s max downsampling rate (e.g. 32 for ResNet [26]¹), and the resolution in each batch should be same. Here the transformation decoder \( D_t \) is responsible for decoding the scale factor \( s \) to make the resolution prediction.

**Noise:** When dealing with the real-world scenarios, the Gaussian noise model is usually adopted to simulate the noises from camera sensor noise [33], low-light noise [60] to quantization noise [50], etc. Therefore, we adopt a zero-mean additive white Gaussian noise (AWGN) model as \( n \) in Eq.1. The variance \( \sigma \) is randomly chosen from a uniform distribution \( U(0, 25/255) \) (e.g. 13.3/255).

4. RestoreDet

Our RestoreDet is built upon CenterNet [73] due to its concise structure. We emphasize that RestoreDet is a generic framework that could also be implemented on

¹A 512 × 512 image \( x \). For example, the degraded image \( t(x) \)'s resolution must be integral multiple of 32, so the down-sampling rate \( s \) is chosen from \([512, 512, 512, 512, ... 512]\).
other popular object detectors [47, 8, 6, 21] and backbones [27, 20].

4.1. CenterNet

CenterNet [73] is an efficient one-stage anchor-free object detector. We show its vanilla structure of [73] in Fig. 2 (a). Input image is fed to the backbone, i.e., ResNet18 [26], to extract $1/32$ bottleneck feature, and then upsampled to a $1/4$ feature map by three $\times2$ deconvolution blocks. This $1/4$ feature map is passed to scaling blocks with three independent convolution blocks to generate the final feature maps. Based on this feature map, there are prediction heads conducting class-wise bbox center detection, bbox height and width regression, offset regression respectively. For more details, please refer to [73].

The CenterNet could be decomposed into an encoder-decoder style structure. Here, we denote the network backbone part (gray part in Fig.2.a) as encoder $E$. The object detection decoder $D_o$, comprised by three prediction heads (colored in orange in Fig.2 (a)), decodes the object information.

4.2. Architecture and Training Pipeline

Fig.2 (b) illustrates how to implement our RestoreDet based on CenterNet [73]. The detailed training procedure is given in Algo.1. When training the RestoreDet, the original HR image $x$ and transformed degraded LR image $t(x)$ are sent to the encoder $E$ to encode the degradation equivariant representation. Here, we directly use the encoder $E$ of CenterNet but duplicate it into a shared weight Siamese structure to receive the HR and LR image, respectively.

With this encoder $E$, the encoded representative $E(x)$ and $E(t(x))$ would have same channel but different size. Then we apply a global average pooling layer $G$ to pool them into the same shape 2-D tensor. $G(E(x))$ and $G(E(t(x)))$ are concatenated together for the transformation decoder $D_t$, which comprises only two FC layers. $D_t$ predicts the transformation parameters $\hat{k}, \hat{\theta}, \hat{n}$, as illustrated in Sec. 3. $k$ represents the Gaussian kernel size, $\theta$ represents the down-sample factor and $n$ represents noise level. While training, all the ground truth $k, n, \theta$ is normalized to $0 \sim 1$ in their own type. For transformation decoder $D_t$, we have MSE loss $l_{trans}$ between $(\hat{k}, \hat{\theta}, \hat{n})$ and $(k, \theta, n)$.

We only put representation $E(t(x))$ of LR image $t(x)$ into upsampling blocks with three $\times2$ deconvolution blocks to generate the final feature map for object detection decoder $D_o$ and arbitrary resolution restoration decoder (ARRD) $D_r$. As illustrated by the red arrows in Fig.2, we introduce the skip connection on $1/8$ and $1/16$ feature maps between the backbone encoder $E$ and the deconvolution blocks. Fusing features from different scales could enhance semantic information and contribute to the subsequent $D_o$ and $D_r$.

We further regularize the representation encode $E$ with our unique ARRD $D_r$ that recovers HR image $\hat{x}$. Since the downsampling rate $s$ is not a fixed integral number in the training stage, ARRD could deal with an arbitrary scale factor. Transformation decoder $D_t$ and ARRD $D_r$ could force the encoder $E$ to not only capture the dynamics of how images change under different transformations, but also extracts the complex patterns of visual structures. Since ARRD $D_r$ aims to recover the original resolution of clean image $x$ from $E(t(x)))$, it could also support the object de-
tection decoder $D_o$ with more detailed features. Motivated by the learnable resizer model [52], we design this decoder with a residual bilinear model shown in Fig. 2, which ends up with a $\times 4$ pixel shuffle layer [51]. ARRD is a light weight structure that uses fewer parameters (0.06M) compared to the backbone encoder (11.17M), upsampling blocks (3.61M) and detection decoder $D_o$ (0.12M). The ARRD loss $l_d$ is defined as an L1 loss between output image $\hat{x}$ and ground truth image $x$.

\[
l_d = |\hat{x} - x|_1 = |D_r(E(t(x))) - x|_1
\]

\[
l_{trans} = ||D_t[E(t(x)), E(x)] - (k, s, n)||_2^2
\]  

(2)

We adopt the three CenterNet prediction heads as the object detection decoder $D_o$ to conduct detection on the final feature map generated by the upsampling block.

As shown in Algo.1, we optimise the total loss $l_{total}$ including detection loss (classwise bbox center loss, bbox width and height loss, bbox offset loss) $l_{obj}$, transformation loss $l_{trans}$ and data restoration loss $l_d$.

\[
l_{total} = l_{obj} + \lambda_1 \cdot l_{trans} + \lambda_2 \cdot l_d,
\]  

where $\lambda_1$ and $\lambda_2$ is the non-negative parameters for loss balancing. They are respectively set to 8 and 0.8 in our experiments.

4.3. Inference Procedure

The inference procedure only involves encoder $E$, upsampling block and object detection decoder $D_o$, as illustrated in Fig.2. Specifically, the encoder $E$ encodes the input target image before $D_o$ performs the detection. Compared to explicitly pre-processing image for high-level tasks [13, 50, 25], our RestoreDet saves much computational time as we avoid reconstructing HR details of data.

We could also reconstruct the HR image with our ARRD decoder $D_r$. Very interestingly, our restored images $\hat{x}$ are more machine vision oriented and exhibit artifacts around the center of the object, as shown in Fig.3.

5. Experiments and Details

5.1. Datasets and Implementation Details

We build our framework based on the open-source object detection toolbox mmdetection [7]. Two well-known object detection datasets, KITTI [22] and MS COCO [39], have been used to evaluate our method. Throughout the experiments, the backbone, ResNet-18 [26], is pre-trained on ImageNet [14] dataset and the training settings adopt the same data augmentation methods (random crop, multi-size, random flip). All input images are resized to $512 \times 512$ in the training stage.

KITTI [22] is a popular small object detection dataset for autonomous driving. For KITTI dataset, we evaluate car class. All of the models are trained on KITTI train set and evaluated on KITTI validation set. They have been trained on a single RTX 6000 GPU for 110 epochs with SGD optimizer. The batch size is set to 16 and the momentum and weight decay are set to 0.9 and 1e-4. The initial learning rate of encoder, detection and data restoration decoder is 1e-3, while the transformation decoder part is 1e-4. The learning rate warms up [44] at first 500 iterations and decays to one-tenth at 100 and 105 epochs. For evaluation, we report the mean average precision (mAP) rate at IoU threshold of 0.5.

MS COCO [39] is a popular large dataset, which contains over 10 million images with 80 categories. In the experiment, we evaluate all 80 classes. For COCO dataset training, all models are trained on COCO 2017 train set and evaluated on COCO 2017 validation set. They have been trained on 4 RTX 6000 GPUs for 140 epochs, with batch size 32. The initial learning rate is the same as KITTI setting and warms up at the first 500 iterations, then decays to one-tenth at 90 and 120 epochs. For evaluation, we report evaluation metric of COCO indexes as: \{AP, AP$_S$, AP$_M$, AP$_L$\} which shows object detection performance on different scales.

In the testing stage, all the results are tested on a single RTX 6000 GPU. We compare the speed by reporting the frames per second (FPS) in the experiments of COCO dataset.

---

**Algorithm 1** RestoreDet Algorithm Pipeline

1. **Data Generation:**
   - B: batch size, C: channel, H: image height, W: image width
   - **inputs:** HR image $x = (B, C, H, W)$, down-sample factor $s \sim (1.0, 4.0)$
   - **outputs:** down-sampling degraded LR image $t(x) = (B, C, H', W')$

2. **Training:**
   - **inputs:** HR image $x = (B, C, H, W)$, and degraded LR image $t(x) = (B, C, H', W')$
   - **outputs:** detection output, estimated SR image $\hat{x}$, estimated transformation $\hat{t}$

3. **Encoding:**
   - $x \rightarrow E_x(t(x)) \rightarrow E(t(x))$
4. **Decoding:**
   - transformation decoding: $\hat{t} = D_t([E(x), E(t(x))])$
   - data restoration decoding: $\hat{x} = D_r(E(t(x)))$
   - detection decoding: detection results $= D_o(E(t(x)))$
Table 1. The detection result on MS COCO [39] and KITTI [22] with randomly degradation (down-sample rate, kernel type and noise level) process. It can be seen that our RestoreDet get superior performance and decent inference speed among various methods. Best result are high lighted.

| model          | train set | Pre-process to test set | KITTI (car) | COCO (80 classes) | FPS (COCO) |
|----------------|-----------|-------------------------|-------------|-------------------|------------|
|                |           |                         | AP$_{50}$   | AP$_{AP_{50}}$    | AP$_{AP_{95}}$ | AP$_{AP_{95}}$ | AP$_{AP_{95}}$ | AP$_{AP_{95}}$ | FPS (COCO) |
| Detection Performance on Original Set (Upper Bound) |           |                         | 85.6        | 30.0              | 10.6       | 0.332       | 0.472       | 51.0   |
| Interpolation Methods |           |                         |             |                   |            |             |             |        |
| CenterNet      | N         | -                       | -           | -                 |            |             |             |        |
| Normal Deep Super Resolution Methods |           |                         |             |                   |            |             |             |        |
| CenterNet      | N         | bicubic                 | x2          | 50.6              | 16.2       | 4.1         | 15.3        | 31.1   | 50.0   |
| Normal Deep Super Resolution Methods |           |                         |             |                   |            |             |             |        |
| CenterNet      | N         | RCAN [69]               | x2          | 47.5              | 13.8       | 3.0         | 13.9        | 26.4   | 50.2   |
| CenterNet      | N         | HAN [45]                | x2          | 46.9              | 14.8       | 3.0         | 14.8        | 27.9   | 49.7   |
| CenterNet      | N         | IMDN [29]               | x2          | 51.3              | 15.0       | 3.5         | 14.3        | 27.4   | 49.6   |
| CenterNet      | N         | PAN [71]                | x2          | 50.5              | 14.8       | 3.3         | 14.1        | 27.8   | 50.1   |
| CenterNet      | N         | SWIN-IR [37]            | x2          | 51.6              | 15.2       | 3.5         | 14.9        | 28.4   | 50.2   |
| Real World Super Resolution Methods |           |                         |             |                   |            |             |             |        |
| CenterNet      | N         | Real-SR [5]             | x2          | 55.6              | 14.2       | 3.1         | 12.4        | 29.5   | 51.1   |
| CenterNet      | N         | RRDB [38]               | x2          | 43.1              | 10.4       | 2.8         | 11.9        | 20.3   | 50.6   |
| CenterNet      | N         | Blind-SR [11]           | x2          | 64.2              | 15.9       | 3.0         | 14.6        | 34.2   | 49.8   |
| CenterNet      | N         | BSRGAN [65]             | x2          | 70.8              | 16.8       | 3.9         | 15.8        | 36.9   | 51.4   |
| Image Restoration Methods |           |                         |             |                   |            |             |             |        |
| CenterNet      | N         | BM3D                    | -           | 50.9              | 10.4       | 0.8         | 6.8         | 27.9   | 86.0   |
| CenterNet      | N         | Cycle-ISP [64]          | -           | 56.7              | 10.6       | 1.3         | 7.1         | 28.4   | 84.5   |
| CenterNet      | N         | IMDN (AS) [29]          | -           | 46.1              | 9.8        | 0.5         | 6.0         | 27.0   | 84.6   |
| CenterNet      | N         | DAGL [9]                | -           | 57.1              | 11.4       | 1.2         | 7.2         | 28.8   | 83.8   |
| Super Resolution Methods + Image Restoration Methods |           |                         |             |                   |            |             |             |        |
| CenterNet      | N         | IMDN+DAGL               | x2          | 60.2              | 15.4       | 3.5         | 15.2        | 27.9   | 49.7   |
| CenterNet      | N         | SWIN-IR+DAGL            | x2          | 59.3              | 15.9       | 3.2         | 15.8        | 28.7   | 49.9   |
| Different Training Schemes |           |                         |             |                   |            |             |             |        |
| CenterNet      | N         | -                       | -           | 42.2              | 14.5       | 1.2         | 10.4        | 38.6   | 87.3   |
| CenterNet      | L         | -                       | -           | 76.0              | 16.4       | 1.5         | 12.4        | 41.3   | 87.4   |
| CenterNet      | N+L       | -                       | -           | 76.6              | 16.5       | 1.8         | 12.2        | 42.2   | 87.3   |
| RestoredDet (w/o $D_r$) | N+L       | -                       | -           | 80.0              | 17.7       | 4.8         | 15.8        | 41.0   | 43.6   |
| Proposed Method |           |                         |             |                   |            |             |             |        |
| RestoreDet (w/o $D_r$) | N+L       | -                       | -           | 79.8              | 17.9       | 2.5         | 15.9        | 42.5   | 87.4   |
| RestoreDet (w D_r) | N+L       | -                       | -           | 80.5              | 18.2       | 3.0         | 16.4        | 43.0   | 87.4   |

5.2. Experiment Results

To synthesize different degrees of degraded images, we generate the LR image with Eq. 1 with random down-sampling rate (random chosen from a uniform distribution $U(1, 4)$), random blur kernel (isotropic Gaussian kernel $k_{iso}$, anisotropic Gaussian kernel $k_{aniso}$ and no blur kernel $k_{none}$), random noise level (AWGN noise with variance randomly chosen from $U(0, 25/255)$).

At first we compare our methods with SOTA image restoration methods, as shown in Table 1. For SR methods, the tested images would be enhanced by interpolation method, SOTA non-blind SR models [45, 71, 29, 69, 37], SOTA blind SR models [5, 58, 65, 11]. We also compare with latest image restoration methods (do not change image resolution) like [64, 9]. From Table 1, we find that enhanced higher resolution (either by interpolate or SR pre-process) improves the detection performance on small objects, but has a negative impact on the large object as well as the inference speed. Super-resolving the image to an excessively resolution will also lead to very undesirable results. Compared to other non-blind SISR methods [29, 71, 45, 37], blind SR methods [11, 65] show better improve in object detection under diverse degradation.

We also train the CenterNet on different training schemes as shown in Table 1. $N$ means the original normal training set. $L$ corresponds to the LR images from the random degradation operations in Sec.3. $N + L$ denotes a mixed of normal and LR images. To compare with the structure like [25], we use ARRD $D_r$ as the pre-processing module for CenterNet and jointly optimized them. The performance is denoted as the $D_r + $ CenterNet in Table 1. Although $D_r + $ performs well on small object detection, it consumes much time in inference. For our RestoreDet’s
### Table 2. Detection results on MS COCO [39] with fixed downsampling rate (2/4). Our RestoreDet achieves superior performance among the same size evaluation compared to alternatives. Best result are high lighted.

| test set | model | train set | Pre-process to test set | COCO | FPS (COCO) |
|----------|-------|-----------|-------------------------|------|-------------|
|          |       |           | method                  | up-resolution | AP | APs | APm | API |            |
| down 2   | CenterNet | N         | bicubic                 | x2   | 17.7 | 1.3 | 13.6 | 45.1 | 106.2     |
|          |          |           | IMDN [29]               |      | 19.5 | 5.3 | 20.1 | 33.1 | 57.2     |
|          |          |           | HAN [43]                |      | 19.7 | 5.4 | 20.6 | 35.1 | 56.8     |
|          |          |           | BSRGAN [65]             |      | 19.2 | 4.8 | 19.1 | 32.9 | 57.0     |
|          |          |           | BM3D                    |      | 20.4 | 5.8 | 20.8 | 37.8 | 56.5     |
|          |          |           | DAGL [9]                |      | 18.0 | 1.2 | 14.3 | 42.8 | 105.6    |
|          |          |           | RestoreDet (w/o $D_t$)   |      | 19.1 | 1.5 | 15.3 | 46.1 | 109.6    |
|          |          |           | RestoreDet (w $D_t$)     |      | 19.1 | 1.6 | 15.5 | 46.0 | 109.8    |
| down 4   | CenterNet | N         | bicubic                 | x2   | 12.9 | 0.8 | 8.7  | 34.4 | 102.2    |
|          |          |           | IMDN [29]               |      | 9.6  | 2.0 | 9.5  | 17.6 | 57.2     |
|          |          |           | Real-SR [5]             |      | 12.7 | 0.7 | 8.5  | 35.1 | 109.2    |
|          |          |           | RRDB [58]               |      | 10.3 | 1.8 | 10.5 | 16.9 | 56.1     |
|          |          |           | BSRGAN [65]             |      | 11.3 | 2.0 | 9.4  | 17.4 | 55.4     |
|          |          |           | Cycle-ISP [64]          |      | 8.9  | 1.9 | 8.4  | 16.2 | 53.2     |
|          |          |           | DAGL [9]                |      | 13.5 | 1.3 | 11.3 | 38.9 | 104.2    |
|          |          |           | RestoreDet (w/o $D_t$)   |      | 11.2 | 2.1 | 10.9 | 19.8 | 58.1     |
|          |          |           | RestoreDet (w $D_t$)     |      | 10.5 | 0.1 | 5.0  | 38.3 | 123.3    |
|          |          |           | BM3D                    |      | 10.7 | 0.1 | 5.2  | 38.7 | 122.1    |
|          |          |           | Cycle-ISP [64]          |      | 13.4 | 1.6 | 11.4 | 33.1 | 45.2     |
|          |          |           | DAGL [9]                |      | 11.7 | 0.4 | 8.9  | 41.0 | 120.7    |
|          |          |           | RestoreDet (w/o $D_t$)   |      | 12.0 | 0.6 | 9.1  | 41.8 | 120.7    |
|          |          |           | RestoreDet (w $D_t$)     |      | 14.3 | 1.5 | 12.6 | 34.9 | 106.0    |

To better verify the effectiveness of our model, similar to [25], we also evaluate our performance on COCO with a fixed down-sampling setting (2 and 4). We need to emphasize that the RestoreDet model here is trained under previous setting. Since the real-world down-sampling condition is unknown, our RestoreDet should be able to automatically adapt to a new setting. As shown in Table 2, our method continuously gets the best result in the same scale evaluation and different training schemes. Our method also keeps a fast computational speed because we avoid reconstructing the HR details.

#### 5.3. Ablation Study

| (resolution: 1∼4) | AP | Recall |
|-------------------|----|--------|
| CenterNet (N)     | 42.2 | 58.8 |
| CenterNet (L)     | 76.0 | 85.7 |
| + $D_t$ (id = 1)  | 77.2 | 87.7 |
| + $D_t$ (id = 2)  | 77.7 | 88.5 |
| + $D_t$ (id = 3)  | 76.8 | 87.0 |
| + $D_t$           | 79.8 | 90.0 |
| + $D_t$ + $D_t$   | **80.5** | **92.2** |

Table 3: Ablation study on KITTI dataset [22]

To evaluate each component in our RestoreDet, we make an ablation study on KITTI dataset [22]. The evaluation metric is AP rate and Recall rate while the training setting...
is the same as the above section. We report the evaluation results in Table 3.

We first trained the vanilla CenterNet on the original/degraded KITTI dataset and evaluated it on the degraded (down 1 ∼ 4) KITTI dataset as CenterNet (N) / CenterNet (L) in Table 3. The augmented degraded training data has greatly improved the detection results. Then we added the transformation decoder $D_t$ and ARRD $D_r$. For transformation decoder $D_t$, we evaluate the variants when $D_t$ is connected to different feature stages in ResNet backbone ($id = 1, 2, 3$ corresponds to the ResNet 2, 3, 4 stage output, respectively). $+D_r$ refers to adding ARRD and different level feature combination process in the network. The final $+D_r + D_t$ is our full RestoreDet architecture, which achieves the best performance.

### 6. Conclusion

In this paper, we propose a novel self-supervised framework, RestoreDet, to handle object detection for degraded low resolution images. To capture the dynamics of feature representations under diverse resolution and degradation conditions, we propose a degradation equivariant representation that is generic and could be implemented on popular detection architectures. To further combine the strength of the existing progress on super resolution (SR), we also introduce an arbitrary-resolution restoration decoder that supervises the latent representation to preserve the visual structure. The extensive experiments demonstrate that our RestoreDet achieves SOTA results on two mainstream public datasets among different degradation conditions (resolution, noise and blur).
References

[1] Mahmoud Afifi and Michael S. Brown. What else can fool deep learning? addressing color constancy errors on deep neural network performance. In International Conference on Computer Vision (ICCV), 2019.

[2] Michal Irani Assaf Shocher, Nadav Cohen. "zero-shot" super-resolution using deep internal learning. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[3] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Finding tiny faces in the wild with generative adversarial network. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 21–30, 2018.

[4] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.

[5] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In Proceedings of the IEEE International Conference on Computer Vision, 2019.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.

[7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhua Chen, and Bernard Ghanem. Super-resolution using deep internal learning. In Proceedings of the IEEE International Conference on Computer Vision, 2019.

[8] Qi. Chen, Yingming Wang, Tong Yang, Xiangyu Zhang, Jian Cheng, and Jian Sun. You only look one-level feature. In IEEE Conference on Computer Vision and Pattern Recognition, 2021.

[9] Mou Chong, Zhang Jian, and Wu Zhuoyuan. Dynamic attention graph learning for image restoration. In IEEE International Conference on Computer Vision, 2021.

[10] Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML’16, page 2990–2999. JMLR.org, 2016.

[11] Victor Cornillère, Abdelaziz Djelouah, Wang Yifan, Olga Sorkine-Hornung, and Christopher Schroers. Blind image super-resolution with spatially variant degradation. ACM Trans. Graph., 38(6), nov 2019.

[12] Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask act with orthogonal tangent regularity for dark object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 2553–2562, October 2021.

[13] Dengxin Dai, Yujian Wang, Yuhua Chen, and Luc Van Gool. Is image super-resolution helpful for other vision tasks? In IEEE Winter Conference on Applications of Computer Vision (WACV), 2016.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[15] Chao Dong, Chen Change Loy, Kaiming He, and et.al. Learning a deep convolutional network for image super-resolution. In Computer Vision – ECCV 2014, pages 184–199, 2014.

[16] Netaloffee Efraf, Daniel Glasser, Alexander Aarts, Boaz Nadler, and Anat Levin. Accurate blur models vs. image priors in single image super-resolution. In 2013 IEEE International Conference on Computer Vision, pages 2832–2839, 2013.

[17] M. Elad and A. Feuer. Restoration of a single super-resolution image from several blurred, noisy, and undersampled measured images. IEEE Transactions on Image Processing, 6(12):1646–1658, 1997.

[18] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrea Zisserman. The pascal visual object classes (voc) challenge. Int. J. Comput. Vision, 88(2):303–338, June 2010.

[19] Ying Fu, Jian Chen, Tao Zhang, and Yonggang Lin. Residual scale attention network for arbitrary scale image super-resolution. Neurocomputing, 427:201–211, 2021.

[20] Peng Gao, Jiasen Lu, Hongsheng Li, Roozbeh Mottaghi, and Aniruddha Kembhavi. Container: Context aggregation network. arXiv preprint arXiv:2106.01401, 2021.

[21] Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast convergence of DETR with spatially modulated co-attention. CoRR, abs/2101.07448, 2021.

[22] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[23] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[24] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[25] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Task-driven super resolution: Object detection in low-resolution images. CoRR, abs/1803.11516, 2018.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

[27] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR, abs/1704.04861, 2017.
[54] Igor Vasiljevic, Ayan Chakrabarti, and Gregory Shakhnarovich. Examining the impact of blur on recognition by convolutional networks, 2017.

[55] Li Wang, Dong Li, Yousong Zhu, Lu Tian, and Yi Shan. Dual super-resolution learning for semantic segmentation. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3773–3782, 2020.

[56] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jundang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10581–10590, June 2021.

[57] Wenjing Wang, Wenhan Yang, and Jiaying Liu. Hla-face: Joint high-low adaptation for low light face detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 16195–16204, 2021.

[58] Xintao Wang, Ke Yu, Shixiang Wu, Jinyin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In Laura Leal-Taixé and Stefan Roth, editors, Computer Vision – ECCV 2018 Workshops, pages 63–79, Cham, 2019. Springer International Publishing.

[59] Zhangyang Wang, Shiyu Chang, Yingzhen Yang, Ding Liu, and Thomas S. Huang. Studying very low resolution recognition using deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[60] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[61] Pengxu Wei, Hannan Lu, Radu Timofte, Liang Lin, and Wangmeng Zuo et al. AIM 2020 challenge on real image super-resolution: Methods and results. CoRR, abs/2009.12072, 2020.

[62] Wenhan Yang, Ye Yuan, Wenqi Ren, and et al. Advancing image understanding in poor visibility environments: A collective benchmark study. IEEE Transactions on Image Processing, 29:5737–5752, 2020.

[63] Xu Yu-Syuan, Tseng Shou-Yao Roy, and et al. Unified dynamic convolutional network for super-resolution with variational degradations. In IEEE Conference on Computer Vision and Pattern Recognition, pages 3217–3226, 2020.

[64] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In CVPR, 2020.

[65] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In IEEE International Conference on Computer Vision, 2021.

[66] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In IEEE Conference on Computer Vision and Pattern Recognition, pages 3217–3226, 2020.

[67] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In IEEE Conference on Computer Vision and Pattern Recognition, pages 3262–3271, 2018.

[68] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2547–2555, 2019.

[69] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In ECCV, 2018.

[70] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2472–2481, 2018.

[71] Hengyuan Zhao, Xiangtao Kong, Jingwen He, Yu Qiao, and Chao Dong. Efficient image super-resolution using pixel attention, 2020.

[72] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 3499–3509. Curran Associates, Inc., 2020.

[73] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. CoRR, abs/1904.07850, 2019.

[74] Wilman W. W. Zou and Pong C. Yuen. Very low resolution face recognition problem. IEEE Transactions on Image Processing, 21(1):327–340, 2012.