- GitHub：https://github.com/dodo0517cc/VRDL_HW4

- Reference：

1. Unsupervised method－ZSSR：https://github.com/assafshocher/ZSSR

   Most super resolution algorithms need to be trained on a specific dataset to obtain the target model. The ZeroShotSR algorithm requires neither prior image samples nor prior training, it uses the internal repetition information of a single image to train a small image-specific CNN during testing.

2. Supervised method－SwinIR：https://github.com/JingyunLiang/SwinIR, https://github.com/cszn/KAIR

   SwinIR is method that use transformer to restore image. The experimental results show that the performance of SwinIR is 0.14-0.45dB higher than the current sota method, and the parameter quantity is also reduced by 67%.

- Brief introduction：

Super Resolution refers to reconstructing a corresponding high-resolution image from an observed low-resolution image by means of software or hardware. Two commonly used indicators for quantitative evaluation of SR quality are PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structure Similarity Index). The higher the two values, the closer the pixel value of the reconstruction result is to the standard. The indicator for evaluation of this homework is PSNR, details are as follows.

The research process of image super-resolution reconstruction based on deep learning is as follows. First, find a set of original images Image1. Second, reduce the resolution of this group of pictures to a group of images Image2, and then reconstruct Image2 super-resolution to Image3 through various neural network structures (Image3 has the same resolution as Image1). Compare Image1 and Image3 through PSNR or other methods, and verify the effect of super-resolution reconstruction. Adjust the node model and parameters in the neural network according to the effect. Eventually, execute the process repeatedly until the result of the fourth step comparison is satisfactory.

PSNR（Peak Signal to Noise Ratio）：

$$MSE = \frac{\sum_{n=1}^{FrameSize} (I_n - P_n)^2}{FrameSize}$$

$$PSNR = 10 \times \log\left(\frac{255^2}{MSE}\right)$$

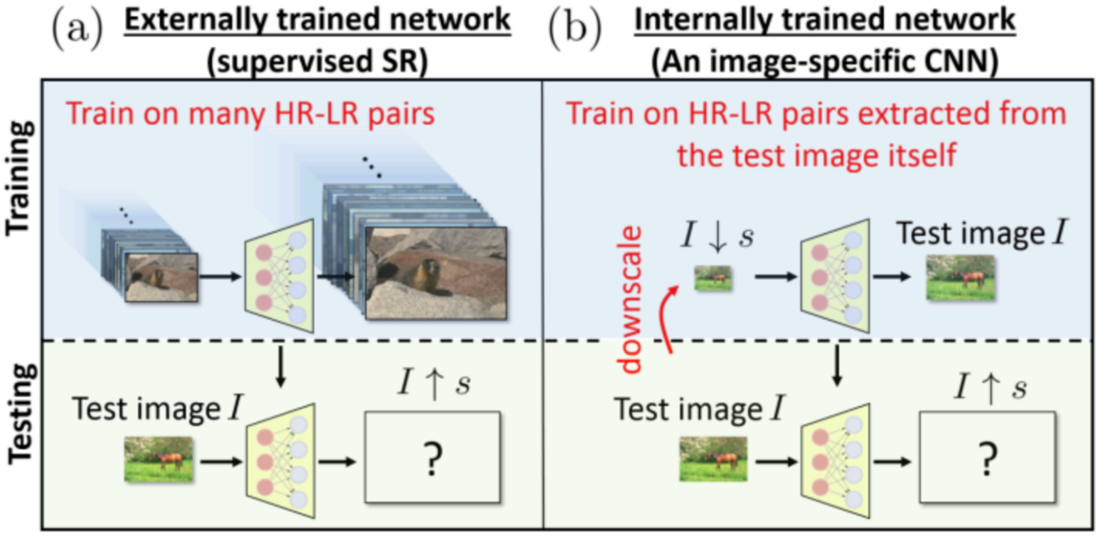
MSE represents the mean square error between the current image I and the reference image P (Mean Square Error). The unit of PSNR is db, and the larger the value, the smaller the distortion.

- Methodology：

✓ Data pre-process：None

✓ Model architecture：

1. ZSSR


(a) Externally trained network (supervised SR)
(b) Internally trained network (An image-specific CNN)
Training
Testing
Train on many HR-LR pairs
Train on HR-LR pairs extracted from the test image itself
downscale
$I \downarrow s$
Test image I
Test image I
$I \uparrow s$
?
Test image I
$I \uparrow s$
?

The right side of the above figure is the training process of ZSSR, and the left side is the training process of other data-driven SR methods. It can be seen that the obvious difference is that ZSSR does not require training data for pre-training, and its training image and testing image are the same images.

Although ZSSR only needs one image I, it needs to use this image I to obtain the label image. First, for the test image I, downsample it by s times to obtain Is_down (down arrow in the figure), and pass CNN reconstructs Is_down into I; in the second step, the test image I is put into CNN again, and the s-fold high-resolution image Is_up will be obtained at this time (the upward arrow in the figure).

At the same time, ZSSR also performs data enhancement work: downsampling different multiples, rotating at different angles, and horizontal and vertical symmetry. From this, multiple HR-LR (high resolution, low resolution) pairs can be obtained and trained on these paired images.

Its main contributions are as follows:

(1) The first CNN super-resolution algorithm built in an unsupervised manner

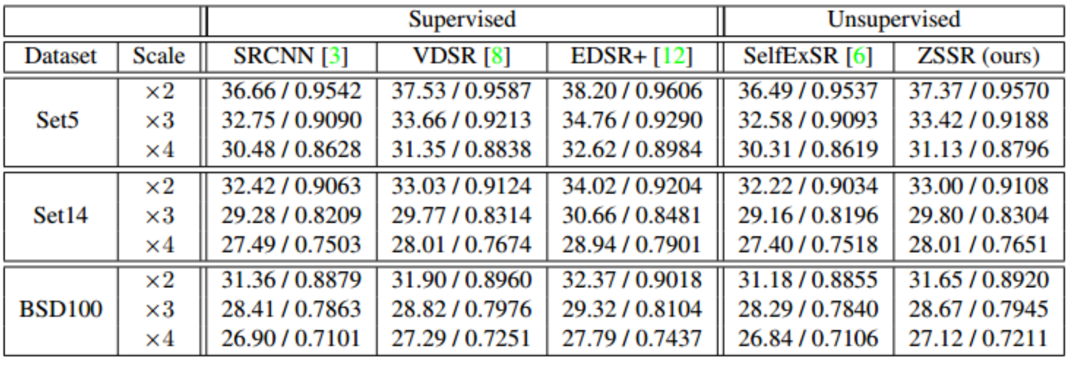
(2) It can process images under non-ideal conditions, such as old historical photos, photos taken by mobile phones and photos from the Internet

(3) Pre-training is not required, and the amount of computation is small

(4) There is no size limit, and it can be applied to SR tasks of any scale

(5) Compared with other pre-training-based methods, ZSSR can achieve comparable results in ideal cases, and better results than pre-training methods in non-ideal cases

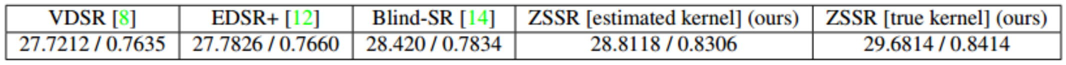
Test in the original paper—Ideal conditions：



		Supervised			Unsupervised	
Dataset	Scale	SRCNN [3]	VDSR [8]	EDSR+ [12]	SelfExSR [6]	ZSSR (ours)
Set5	×2	36.66 / 0.9542	37.53 / 0.9587	38.20 / 0.9606	36.49 / 0.9537	37.37 / 0.9570
	×3	32.75 / 0.9090	33.66 / 0.9213	34.76 / 0.9290	32.58 / 0.9093	33.42 / 0.9188
	×4	30.48 / 0.8628	31.35 / 0.8838	32.62 / 0.8984	30.31 / 0.8619	31.13 / 0.8796
Set14	×2	32.42 / 0.9063	33.03 / 0.9124	34.02 / 0.9204	32.22 / 0.9034	33.00 / 0.9108
	×3	29.28 / 0.8209	29.77 / 0.8314	30.66 / 0.8481	29.16 / 0.8196	29.80 / 0.8304
	×4	27.49 / 0.7503	28.01 / 0.7674	28.94 / 0.7901	27.40 / 0.7518	28.01 / 0.7651
BSD100	×2	31.36 / 0.8879	31.90 / 0.8960	32.37 / 0.9018	31.18 / 0.8855	31.65 / 0.8920
	×3	28.41 / 0.7863	28.82 / 0.7976	29.32 / 0.8104	28.29 / 0.7840	28.67 / 0.7945
	×4	26.90 / 0.7101	27.29 / 0.7251	27.79 / 0.7437	26.84 / 0.7106	27.12 / 0.7211

Table 1: **Comparison of SR results for the 'ideal' case (bicubic downscaling).**

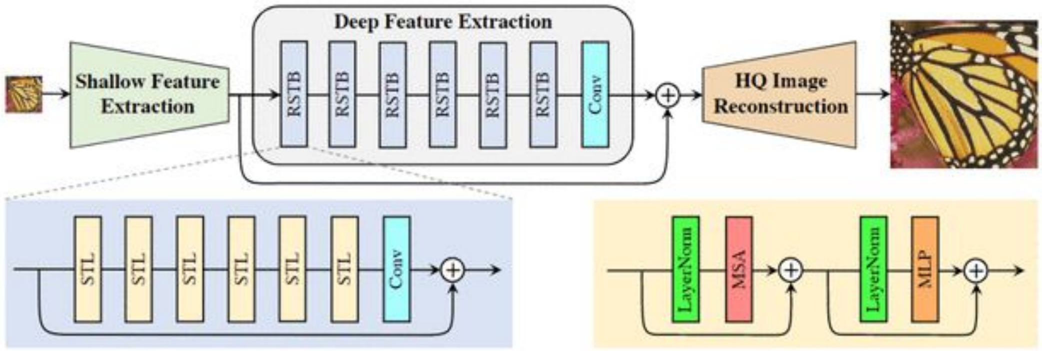
Test in the original paper—Non-ideal conditions：



VDSR [8]	EDSR+ [12]	Blind-SR [14]	ZSSR [estimated kernel] (ours)	ZSSR [true kernel] (ours)
27.7212 / 0.7635	27.7826 / 0.7660	28.420 / 0.7834	28.8118 / 0.8306	29.6814 / 0.8414

Table 2: **SR in the presence of unknown downscaling kernels.** *LR images were generated from the BSD100 dataset using* random *downscaling kernels (of reasonable size). SR×2 was then applied to those images. Please see text for more details.*

2. SwinIR



(a) Residual Swin Transformer Block (RSTB)

(b) Swin Transformer Layer (STL)

It mainly includes three parts: shallow feature extraction, deep feature extraction and high-quality image reconstruction.

(1) Shallow feature extraction：

The shallow feature extraction module uses convolutional layers to extract shallow features, and directly transfers the shallow features to the reconstruction module to preserve low-frequency information.

(2) Deep feature extraction：

The deep feature extraction module is mainly composed of residual Swin Transformer Block (RSTB), each block utilizes multiple Swin Transformer layers (STL) for local attention and cross-window interaction. In addition, a convolutional layer is added at the end of the block to enhance features, and residual connections are used to provide shortcuts for feature aggregation, that is, RSTB consists of multiple STLs and a convolutional layer together to form a residual block,

(3) High-quality image reconstruction：

Combines shallow and deep features to restore high-quality images

✓ Hyperparameters / Configs：

1. SwinIR：

Loss function—L1 loss
Loss function weight—1.2
Learning rate—0.0002
Optimizer—Adam
HR image resize—96
Upsampler—pixelshuffle
Batch size—16
LR patch size—48

2. ZSSR：

Learning rate—0.001
variance of weight initializations—0.1
augment_leave_as_is_probability—0.05
augment_no_interpolate_probability—0.45
augment_min_scale—0.5
augment_scale_diff_sigma—0.25
augment_shear_sigma—0.1
augment_allow_rotation—True
Random crop size—128
Downscale method—cubic
Upscale method—cubic

- Summary

I’ve tried two methods. One is supervised model, SwinIR, and the other is unsupervised model, ZSSR. The SwinIR model removes severe noise interference and preserves high frequency image details for sharper edges and more natural textures. The ZSSR model also reconstruct clear image, but we can see that it’s still blurrier than the SwinIR model. The testing set of this homework is somehow ideal images. That is to say, compared with other pre-training-based methods, ZSSR can really achieve comparable results in ideal cases, and better results than pre-training methods in non-ideal cases.
The final testing PSNR of ZSSR model is roughly 27.8, and the testing PSNR of SwinIR model is roughly 28.2.

Restoration of SwinIR

Restoration of ZSSR