• **Motivation**
  - Widely used bilinear interpolation, cubic interpolation, etc. are handcrafted, and only consider relative distance in defining the interpolation coefficients.
  - Handcrafted interpolation losses much information in downsizing image, which reduces the network performance.

• **Contributions**
  - We propose a new interpolation method DownsizeNet aiming to preserve image information in image downsizing with a sub-network interpolation module at the image preprocessing stage. The interpolation sub-network can be easily used in various vision tasks, and has good generalization ability to different pipelines by making the downsizing result retain a real image.
  - We introduce texture feature into CNN based interpolation coefficient inference. A special floating type pooling layer is designed to spatially align the CNN texture feature and the encoded relative distance map, and this facilitates the pixel-wise interpolation coefficient inference.
  - Experiments on seven pipelines in detection and segmentation tasks demonstrate that our method consistently reduces accuracy drop than widely used bilinear interpolation. Besides, we have demonstrated that our method also outperforms texture only based interpolation, cubic interpolation and area interpolation.

• **Network architecture**
  - Given the high resolution image, and the high resolution H×W and low resolution value h×w, output a low resolution image.
  - The interpolation network is simple, and contains only 2×3 3×3 convolution layer and 1 1×1 convolution layer. Besides, two efficient special layers are proposed. This makes the network efficient.

• **Two special layers**
  - Floating pooling layer. Given a projected point Q’ on high resolution image, the feature vector of P1 is selected as the feature of Q’.
  - Interpolation layer. The last sigmoid can be replaced by softmax, then the interpolation operation is more convenient.

\[
\hat{q} = [q_1, \ldots, q_4|a_1, \ldots, a_4]^T,
\]

\[
s.t. \quad \sum_{i} a_i = 1 \quad 0 \leq a_i \leq 1.
\]

• **Experimental results**
  - Table 1: The performance (mAP) comparison between Bilinear and ours on four detection pipelines. RetinDet is implemented with Caffe, and other networks are implemented with PyTorch. * ~ 300 denotes that one side of the image is resized to 300 while the height and width ratio is fixed.
  - Table 2: The performance (mean IoU) comparison between Bilinear and ours on three segmentation pipelines. FCN is implemented with Caffe, and other networks are implemented with PyTorch. SPP: Spatial Pyramid Pooling. For DeepLabv3+, the image is resized to 0.35 times of the original resolution. For SPNet, the image is resized to 0.3 times of the high resolution. The resolutions 513 and 768 are used in the original codes of DeepLabv3+ and SPNet respectively.
  - Table 3: Test on LightHead about the generalization ability of the pretrained DownsizeNet model on Pascal VOC2007 (mAP metric). The interpolation sub-network is pretrained on DeepLabv3+ for 100, 200 and 300 epochs separately. The LightHead is re-trained with 10 epochs. The image is resized to 0.3 times of the high image resolution for both LightHead and DeepLabv3+.

• **Discussion**
  - DownsizeNet can also be used in feature map resizing (downsizing and upscaling).