SEOUL NATIONAL UNIV. VISION & LEARNING

Better to Follow, Follow to Be Better:
Towards Precise Supervision of Feature Super-Resolution for Small Object Detection

Junhyug Noh Wonho Bae Wonhee Lee Jinhwan Seo Gunhee Kim

Project Page: http://vision.snu.ac.kr/projects/better-to-follow

ICCV 2019 Seoul, Korea

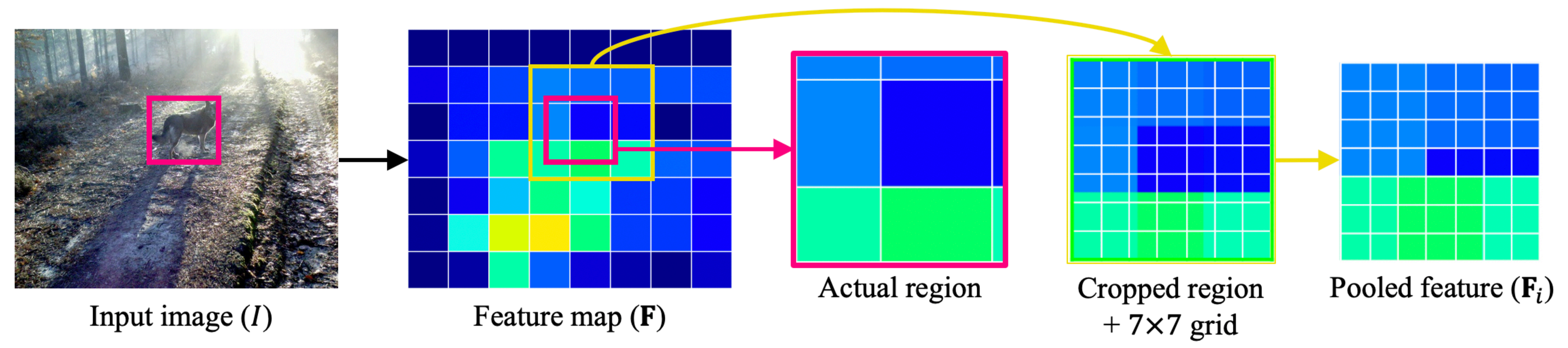
Summary

Problem: poor performance of a proposal-based detector using feature-level super-resolution on small objects

Cause: absence of direct supervision from target features

Solution: novel approach to *"properly"* extract target features as direct supervision

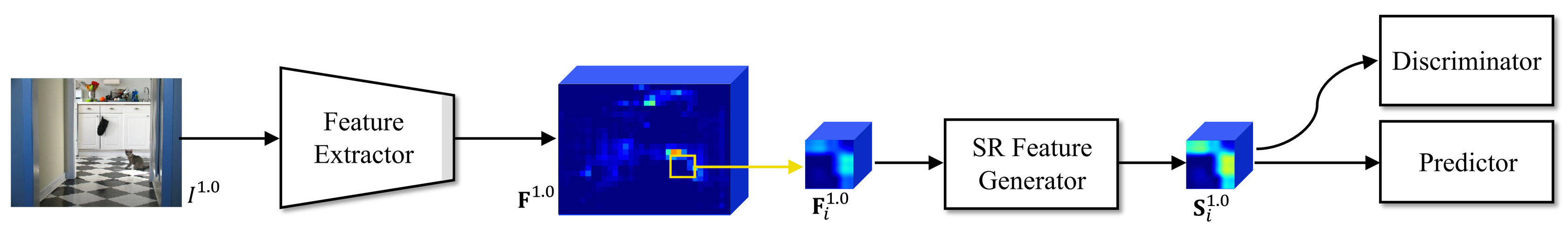
Difficulty of Detecting Small objects



1. RoI pooled features do not contain detailed information due to its size
2. In the process of RoI pooling, internal positions are distorted

Then? **Super-resolve features as large objects!**

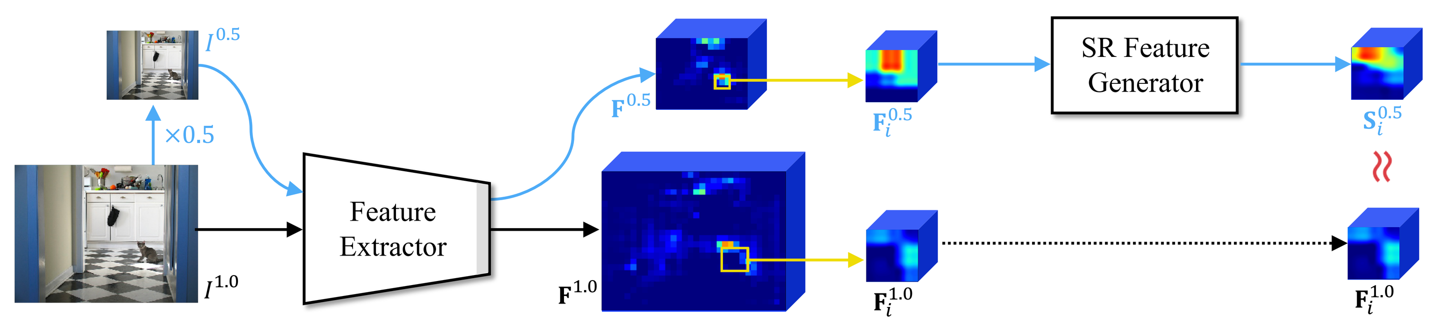
Methods to Generate Super-Resolution



Step 1. Generate super-resolution features ($\mathbf{S}_i^{1.0}$) from an original image

1. To be similar to high-resolution features of a large object (**Discriminator**)
2. So that a class and box offsets of the small object are correctly predicted (**Predictor**)

→ But ***without supervision***, training of the super-resolution feature generator can be unstable



Step 2. Generate super-resolution features ($\mathbf{S}_i^{0.5}$) from a downsampled image

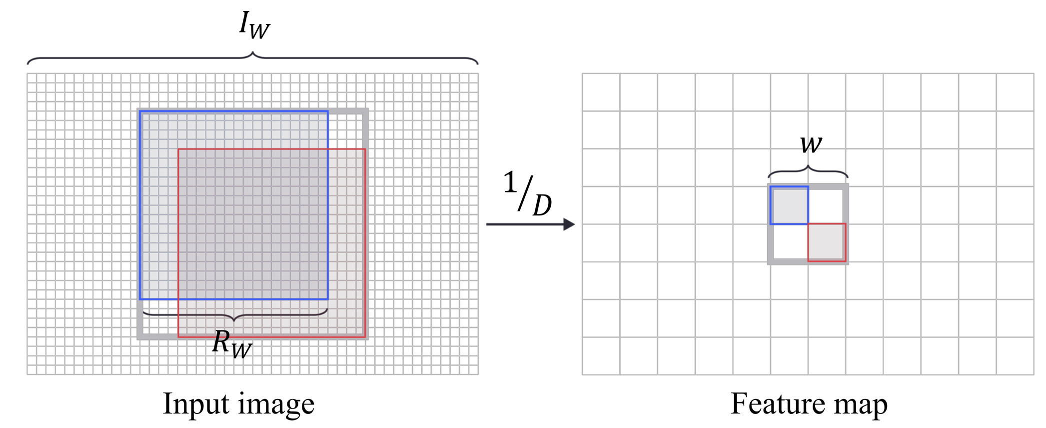
3. To be similar to the corresponding naïve targets ($\mathbf{F}_i^{1.0}$)

→ Even ***with naïve supervision***, it is hard to imitate target features due to high disparity between input ($\mathbf{F}_i^{0.5}$) and target features ($\mathbf{F}_i^{1.0}$)

References

[1] Zhe Zhu, et al. ***Traffic-Sign Detection and Classification in the Wild.*** In CVPR, 2016.
[2] Jianan Li, et al. ***Perceptual Generative Adversarial Networks for Small Object Detection.*** In CVPR, 2017.
[3] Zhenwen Liang, et al. ***Small Object Detection Using Deep Feature Pyramid Networks.*** In Pacific Rim Conference on Multimedia, 2018.
[4] Zibo Meng, et al. ***Detecting Small Signs from Large Images.*** In IRI, 2017.

Mismatch of Relative Receptive Fields

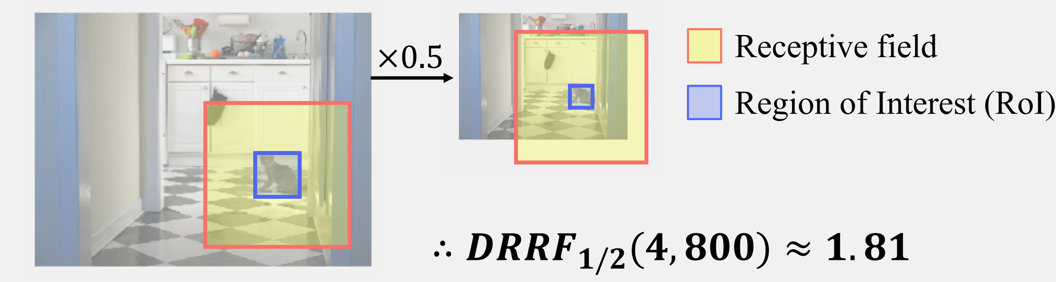


- ***ARF***: Absolute Receptive Field
  $ARF(w) = R_W + (w-1)\times D$
- ***RRF***: Relative Receptive Field
  $RRF(w, I_W) = (R_W + (w-1)\times D)/I_W$
- ***DRRF***: Discrepancy in RRF of the RoIs between the original and downsampled images

$$DRRF_{1/2}(w, I_W) = \frac{RRF(w/2, I_W/2)}{RRF(w, I_W)} = 2 - \frac{w}{c+w}$$

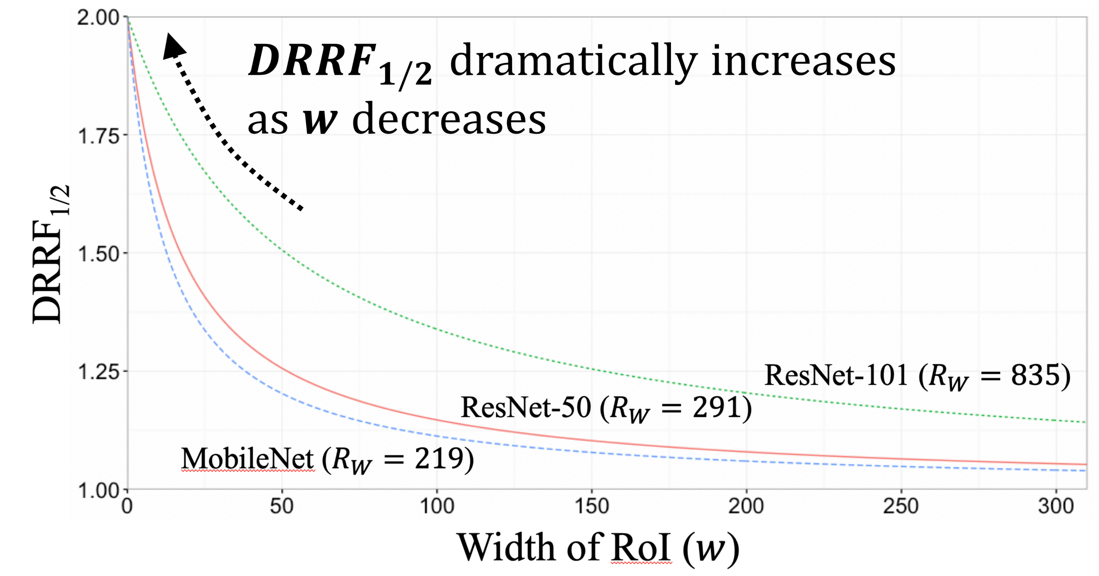
where $c = \frac{R_W}{D} - 1$

e.g. FRCNN ($D = 16$) with ResNet-50 ($R_W = 291$)



$\therefore$ $\boldsymbol{DRRF_{1/2}(4{,}800) \approx 1.81}$

→ RRF of the RoI from the downsampled image is around **1.81 times larger** than that from the original image



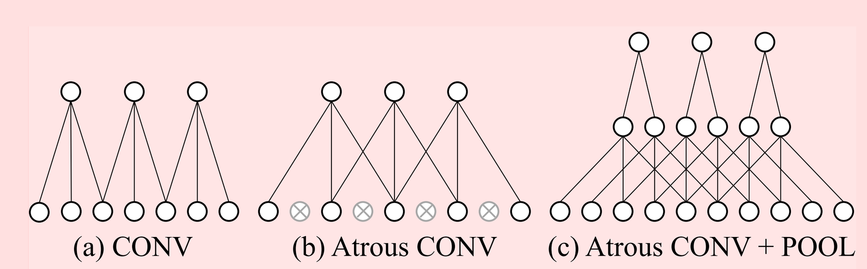
Our Approach

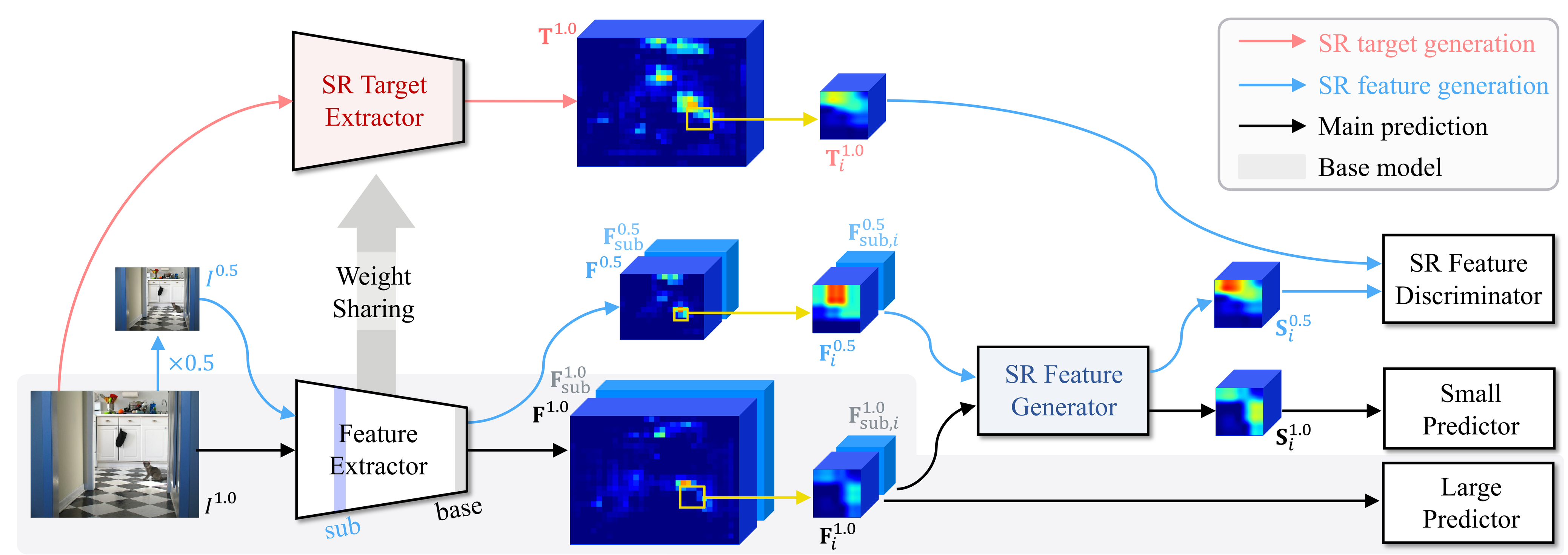
Super-Resolution Target Extractor

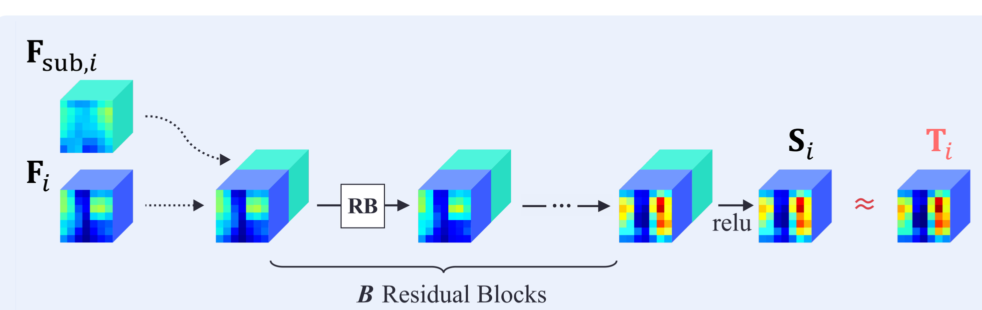
Replace every layer that increases relative field of feature extractor with a layer that doubles it

(k: kernel size, s: stride, r: dilation rate)

- $k\times k$ POOL ($k > 1$) → $2k\times 2k$ POOL → **to use the same weights**
- $k\times k$ CONV ($k > 1, s = 1$) → $\boldsymbol{k\times k}$ **Atrous CONV** ($r = 2, s = 1$)
- $k\times k$ CONV ($k > 1, s = 2$) → $k\times k$ Atrous CONV ($r = 2, \boldsymbol{s = 1}$) + **2×2 POOL** ($\boldsymbol{s = 2}$) → **not to skip every other pixel**







Super-Resolution Feature Generator

Transform $\mathbf{F}_i$ into $\mathbf{S}_i$ with additional input $\mathbf{F}_{sub,i}$ by iteratively refining via B residual blocks

Quantitative Results

Tsinghua-Tencent 100K

1. Results on different backbones (input: 1600×1600)

Model	Small			Medium			Large			Overall		
	Rec.	Acc.	F1	Rec.	Acc.	F1	Rec.	Acc.	F1	Rec.	Acc.	F1
MobileNet	56.1	72.9	63.4	85.1	**84.3**	84.7	90.9	**83.6**	**87.1**	74.7	80.7	77.5
+ Ours	**62.7**	**81.7**	**71.0**	**87.6**	84.0	**85.7**	**91.5**	82.1	86.5	**78.5**	**83.1**	**80.7**
ResNet-50	68.8	81.9	74.9	90.8	93.1	91.9	91.6	92.3	91.9	82.5	89.2	85.7
+ Ours	**78.2**	**86.5**	**82.2**	**94.7**	**93.8**	**94.3**	**93.6**	**93.0**	**93.3**	**88.4**	**91.1**	**89.7**
ResNet-101	69.8	81.5	75.2	90.9	93.5	92.2	92.4	92.0	92.2	83.1	**89.2**	86.0
+ Ours	**86.6**	**82.1**	**84.3**	**95.5**	**93.7**	**94.6**	**93.7**	**92.7**	**93.2**	**91.9**	89.1	**90.5**

- Consistent improvement over the base models regardless of the backbones
- Performance (F1) improvement: small > medium > large

2. Comparison with SOTA models (input: 2048×2048)

Model	Small			Medium			Large			Overall		
	Rec.	Acc.	F1	Rec.	Acc.	F1	Rec.	Acc.	F1	Rec.	Acc.	F1
Zhu *et al.* [1]	87.0	82.0	84.4	94.0	91.0	92.5	88.0	91.0	89.5	–	–	–
Perceptual GAN [2]	89.0	84.0	86.4	96.0	91.0	93.4	89.0	91.0	89.9	–	–	–
Liang *et al.* [3]	**93.0**	84.0	88.3	97.0	**95.0**	95.9	92.0	**96.0**	93.9	–	–	–
SOS-CNN [4]	–	–	–	–	–	–	–	–	–	93.0	90.0	91.5
FRCNN + ResNet-101	80.3	81.6	80.9	94.5	94.8	94.7	94.3	92.6	93.5	89.1	89.7	89.4
+ Ours	92.6	**84.9**	**88.6**	**97.5**	94.5	**96.0**	**97.5**	93.3	**95.4**	**95.7**	**90.6**	**93.1**

3. Comparison of super-resolution methods

Model	Small	Medium	Large	Overall
Base model (ResNet-50)	74.9	91.9	91.9	85.7
+ SR (w.o. supervision)	76.8	93.6	**93.3**	87.5
+ SR (Naïve supervision)	74.4	91.8	92.3	85.3
+ SR (Ours)	**82.2**	**94.3**	**93.3**	**89.7**

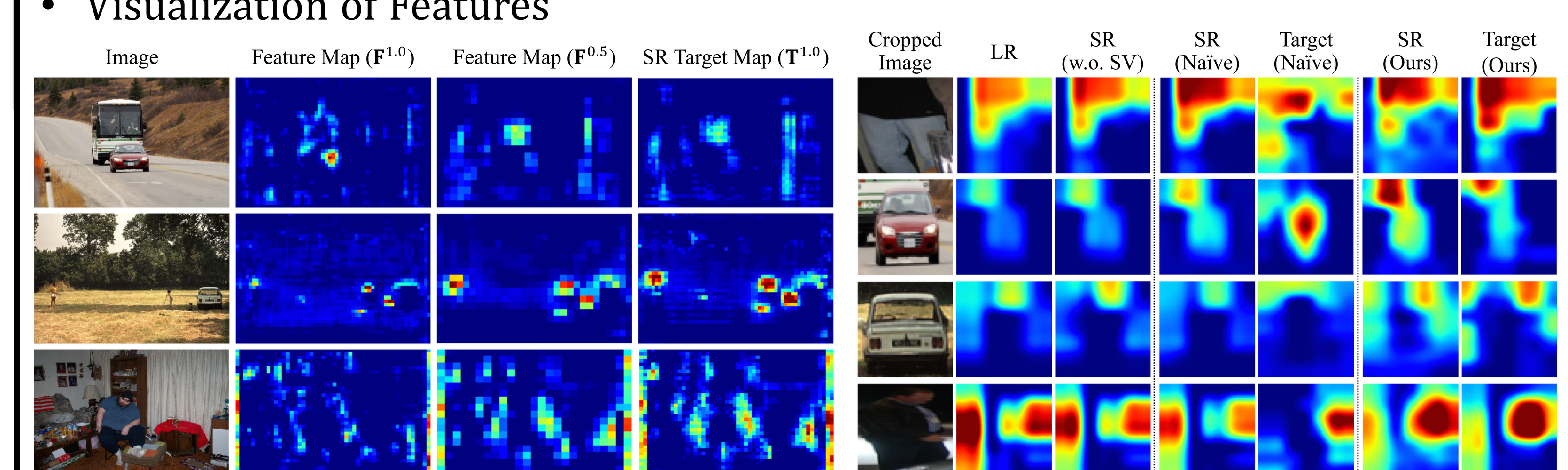
- Performance (F1) improvement: ours > w.o. supervision > naïve supervision
- SR with naïve supervision performs even worse than the base model

PASCAL VOC & MS COCO

Model	PASCAL VOC				MS COCO					
	AP-.5	AP-S	AP-M	AP-L	AP-.5:.95	AP-.5	AP-.75	AP-S	AP-M	AP-L
MobileNet	73.2	5.1	39.3	**76.9**	19.3	38.7	16.9	5.4	20.6	**29.2**
+ Ours	**77.0**	**10.1**	**47.2**	**76.9**	**21.9**	**41.0**	**21.0**	**10.9**	**23.8**	29.0
ResNet-50	77.1	6.8	42.9	81.1	29.5	52.0	29.8	10.2	31.5	**44.7**
+ Ours	**79.1**	**10.5**	**47.9**	**81.4**	**31.2**	**54.2**	**32.4**	**14.3**	**32.4**	**44.7**
ResNet-101	78.8	5.9	46.2	82.3	32.0	54.7	32.8	11.3	34.3	**48.1**
+ Ours	**80.6**	**11.1**	**48.9**	**82.7**	**34.2**	**57.2**	**36.1**	**16.2**	**35.7**	**48.1**

Qualitative Results

- Visualization of Features



Comparison of **features from different extractors** Comparison of **different super-resolution methods**

- Detection results on **Tsinghua-Tencent 100K** (G: TP, R: FP, B: FN)