Unsupervised Sound Source Localization From Audio-Image Pairs Using Input Gradient Map

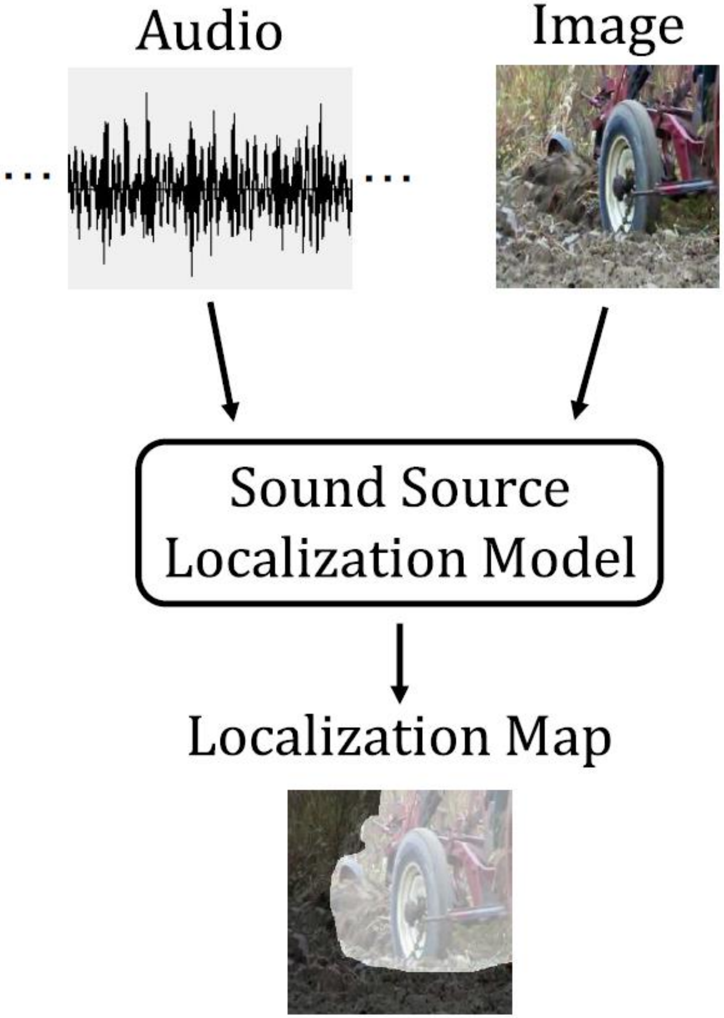
Tomohiro Tanaka, Takahiro Shinozaki

Tokyo Institute of Technology

Background

- Unsupervised Sound Source Localization
    - Input <$\boldsymbol{a}$, $\boldsymbol{v}$>: audio-image pair
    - Output $\widehat{\boldsymbol{y}}$: pixel-wise prediction of sound source location

- It builds a foundation of a self-sustaining intelligent robot that works in the real world



Conventional Methods

- Designed neuron layer based approach
  - NN predicts sound from image [Zhao+, ECCV, 2018]
  - NN evaluates correspondence between sound and image
    - Attention based method [Senocak+, CVPR, 2018]
    - K-means based method [Hu+, CVPR, 2019]
    - CAM based method [Owens+, ECCV, 2018]

✗ Decreased resolution

- Input based approach
  - Occlusion sensitivity method
    [Ephrat+, SIGGRAPH, 2018]
    [Gao+, CVPR, 2019]

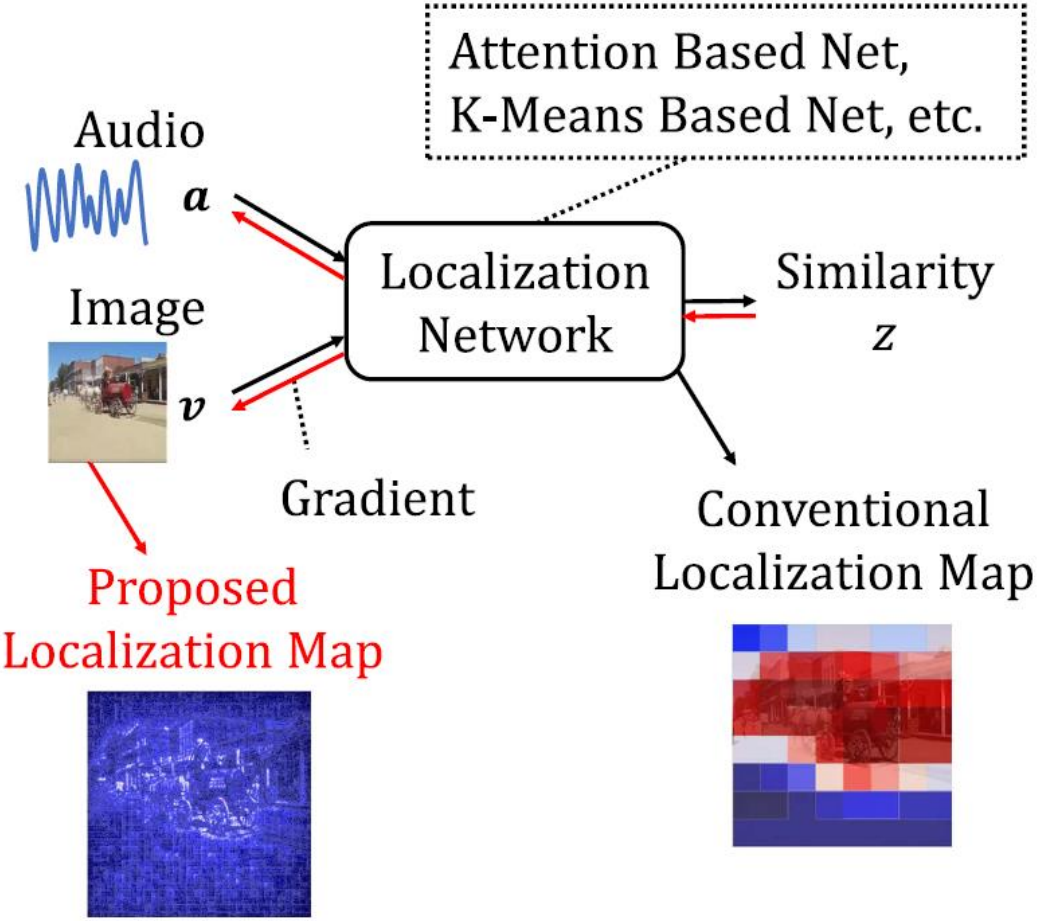
✓ Preserved resolution

✗ High computational cost

Proposed Method

Use input gradient of predicted similarity as localization map

- ✓ High resolution
- ✓ Low computational cost
- ✓ Free network structure

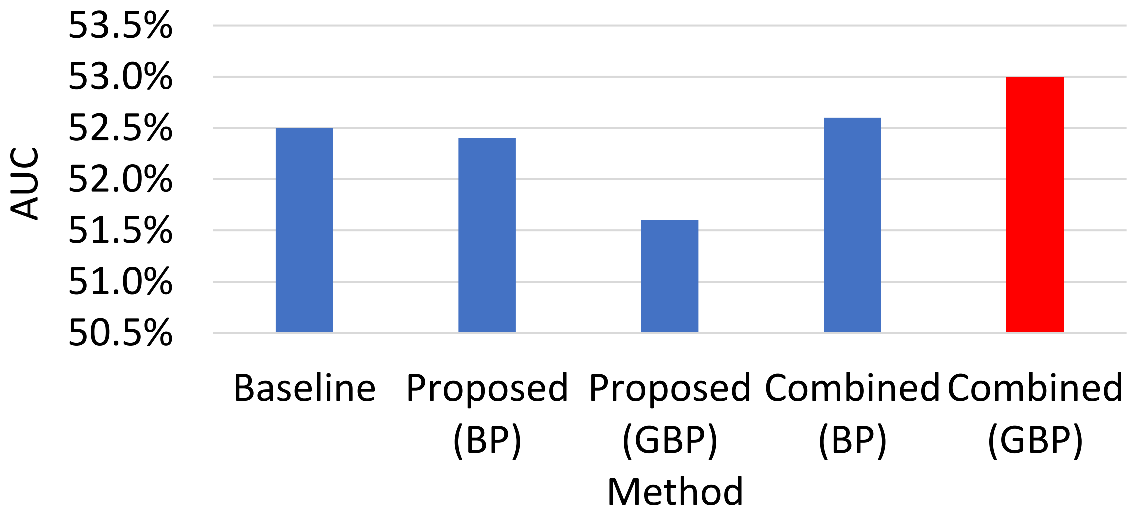


Experimental Setup

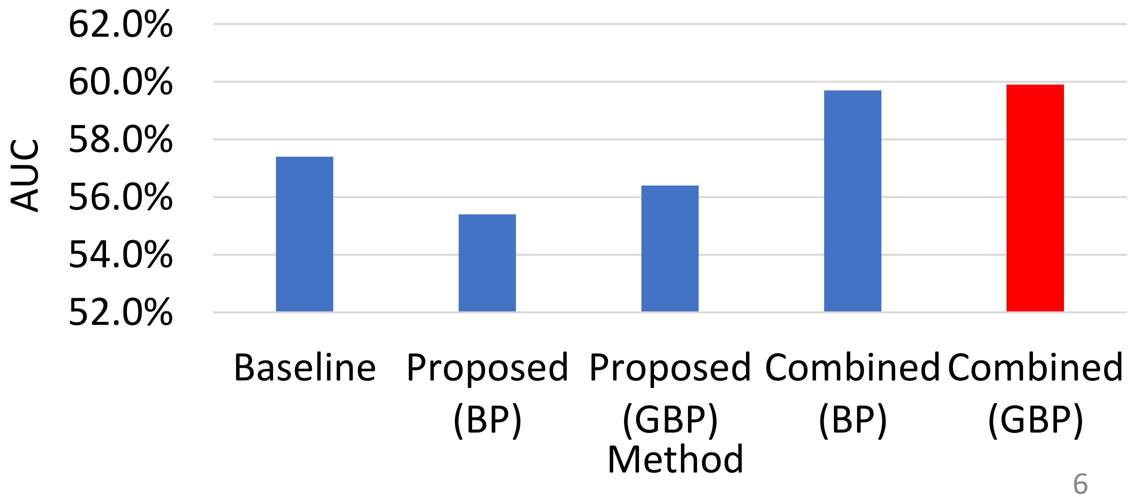
- Dataset: Flickr-SoundNet Dataset
- Baseline:
  - Attention based method (144k training samples) [Senocak+, CVPR, 2018]
  - K-means based method (400k training samples) [Hu+, CVPR, 2019]
- Proposed method:
  - Gradient: Back propagation, Guided back propagation
- Combined method:
  - Combined the baseline and the proposed method by averaging the localization map
- Evaluation:
  - Data: 250 labeled samples
  - Post process: bounding box
  - Metrics: AUC of cIoU [Senocak+, CVPR, 2018]

Unsupervised Experimental Result

Attention based model, 144k Unsup. data

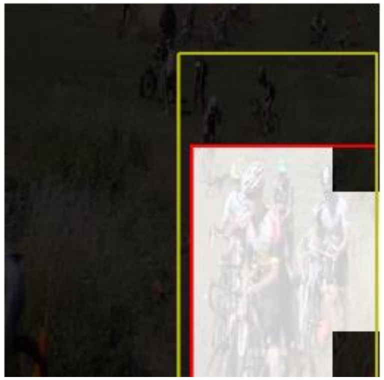


K-means based model, 400k Unsup. data

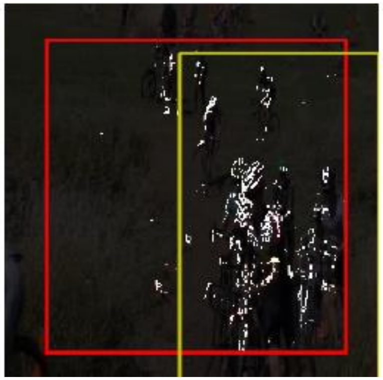


Localization Map Example

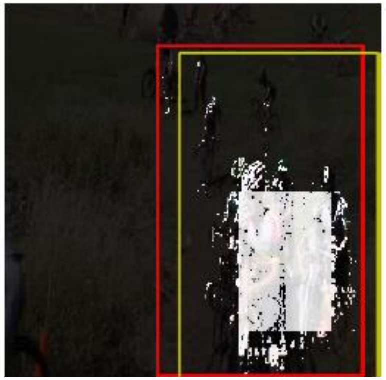
Original



Baseline

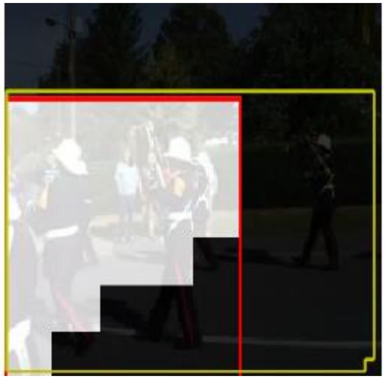


Proposed

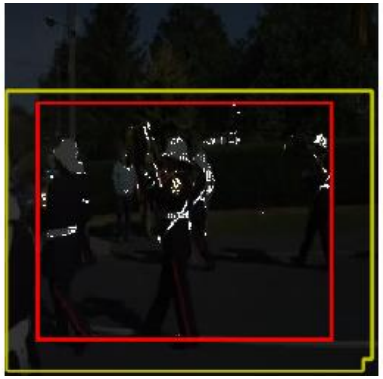


Combined

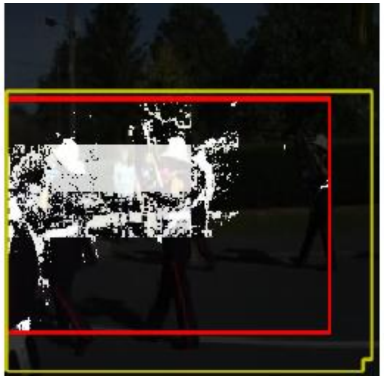
Original



Baseline



Proposed



Combined

Summary

- Proposed an input gradient-based unsupervised sound source localization method
    - low computational cost
    - high resolution
    - Free network structure
- Consistent improvement from the baselines when the conventional and proposed methods were combined
- Future work
    - Extension to video input
    - Evaluation with multiple sound sources