Learning to Localize Sound Source in Visual Scenes

Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, In So Kweon. Senocak,

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

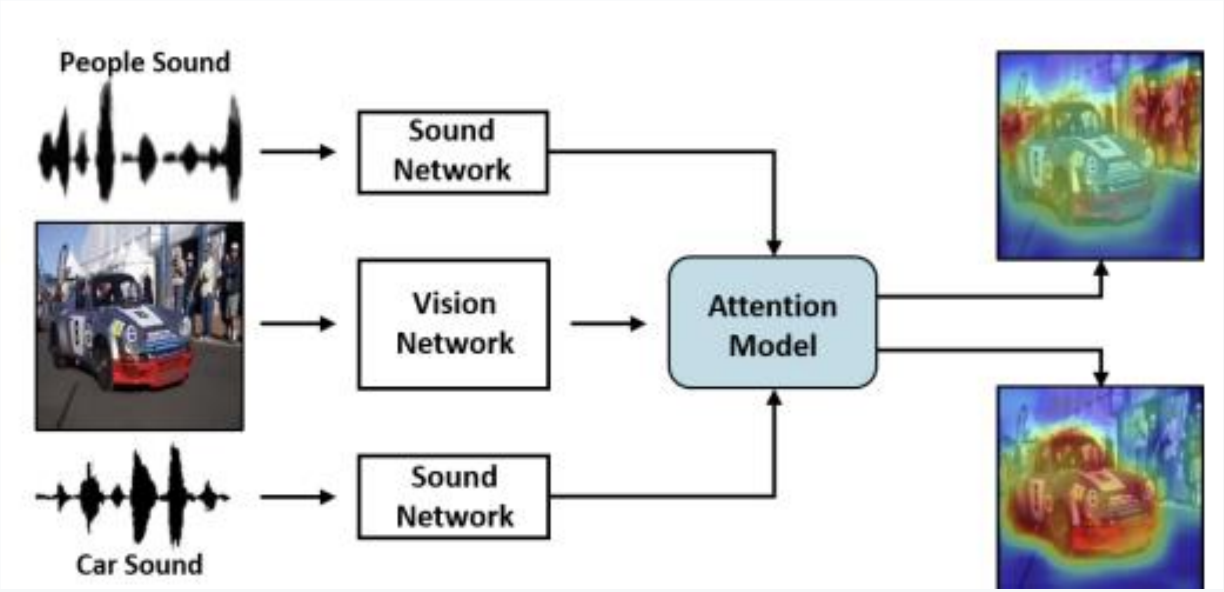
Presenter :
Energy AI – Sohee Kim

◆ Introduction

Human

Car → engine sound

❖ Learning based sound source localization



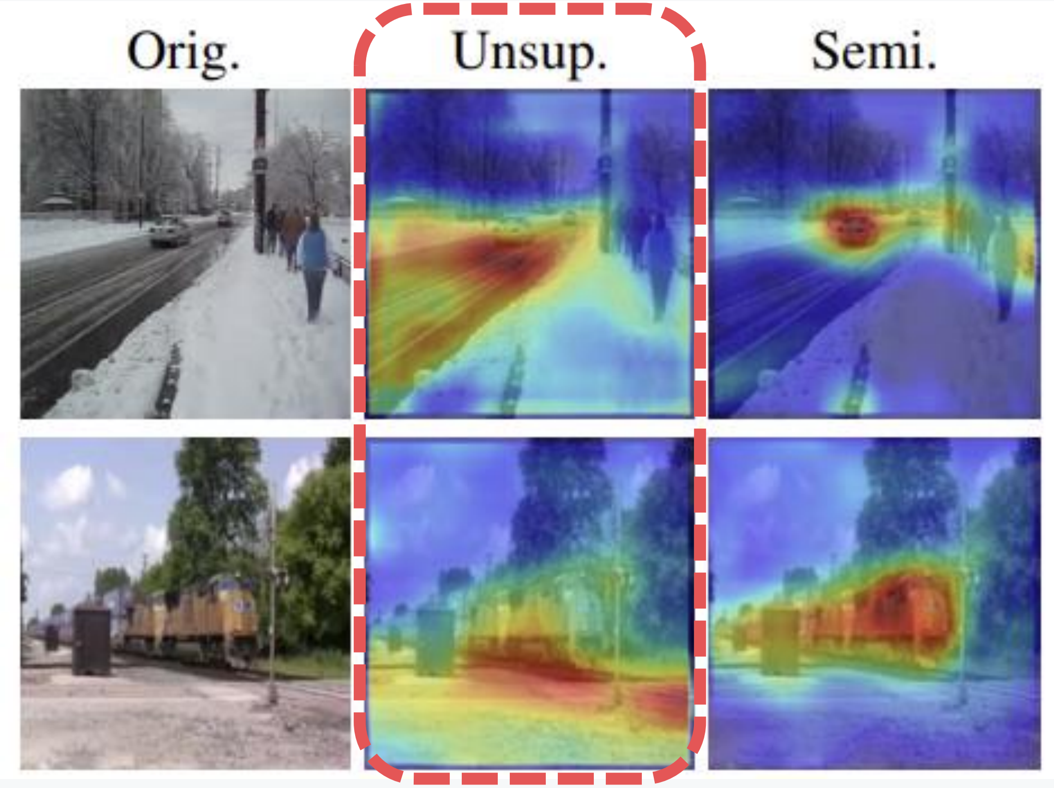
- How to learn to localize the sources (objects) from the sound signals
  - Based on simply watching and listening

◆ Introduction

✓ Learning task from **unlabeled** data = challenging

- Unconstrained video contains
  - Unrelated audio
  - Audio source that is off the screen
- Pigeon superstition

⇒ **Biases** the resulting localization to be unmatched

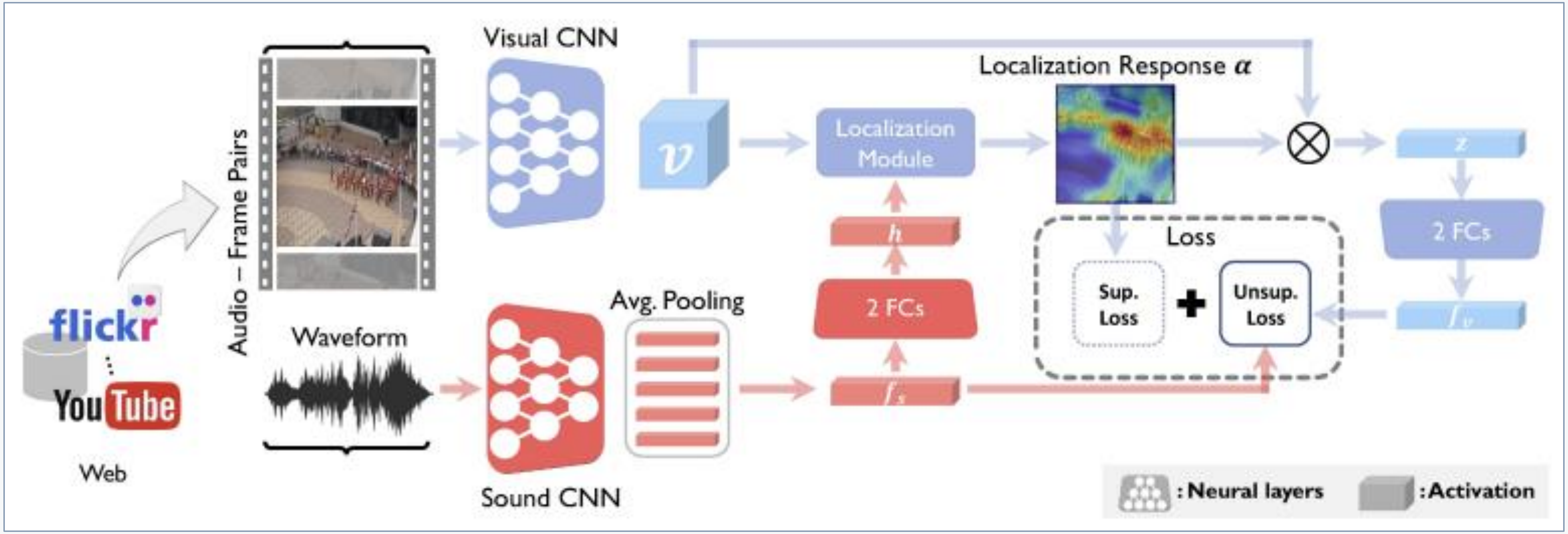


❖ **Semi-supervised** setting

- Provide some prior knowledge
- Add supervised loss
- Annotated data

◆ Introduction

❖ Network Architecture



Introduction

❖ The main contributions

- A learning framework to localize sound source using the attention mechanism, guided by sound information
- A unified end-to-end deep convolutional neural network architecture
  - unsupervised, semi-supervised, fully-supervised
- Collect and annotate a new sound source localization dataset

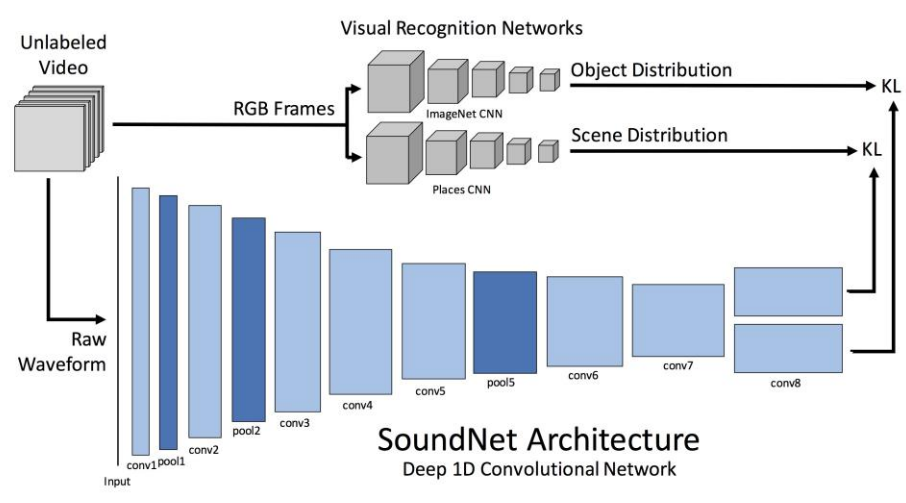
◆ Related Work and Problem Context

❖ Recent work

✓ **SoundNet**

- Visual imagery as supervision for sound

⇒ Audio module in this work



✓ *Aytar et al.* and *Arandjelovic et al.*

- Aligned cross-modal representations
- Activation maps that localize object

  → Not interactively estimated according to the given sound

⇔ **A bridge layer**

- Reveals the localization information of the sound sources

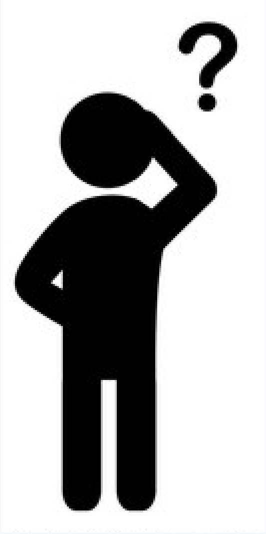
◆ Related Work and Problem Context

❖ Recent work

✓ **Sound source localization capability of humans**

⇒ Guided by visual information

⇒ Two sources of information are closely correlated that humans can unconsciously learn the capability.



◆ Related Work and Problem Context

❖ Recent work

✓ Deep learning methods rely on

- Synchrony of low-level features of sounds and videos
- Spatial sparsity prior of audio-visual events
- Low-dimensionality
- Hand-crafted motion

⇔ An **unsupervised** manner by only watching and listening to videos without using any manually designed constraints such as motion.

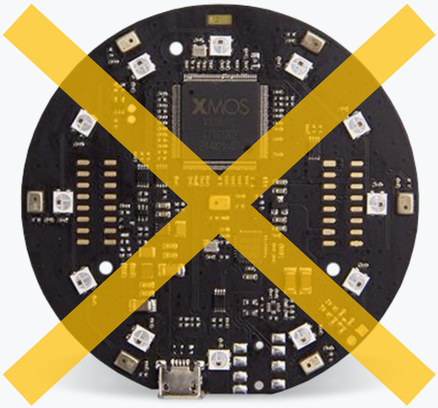
◆ Related Work and Problem Context

❖ Recent work

✓ Acoustic based approach - in surveillance and instrumentation engineering

- Requires specific devices (microphone arrays)

⇒ Without any special devices but a microphone to capture sound

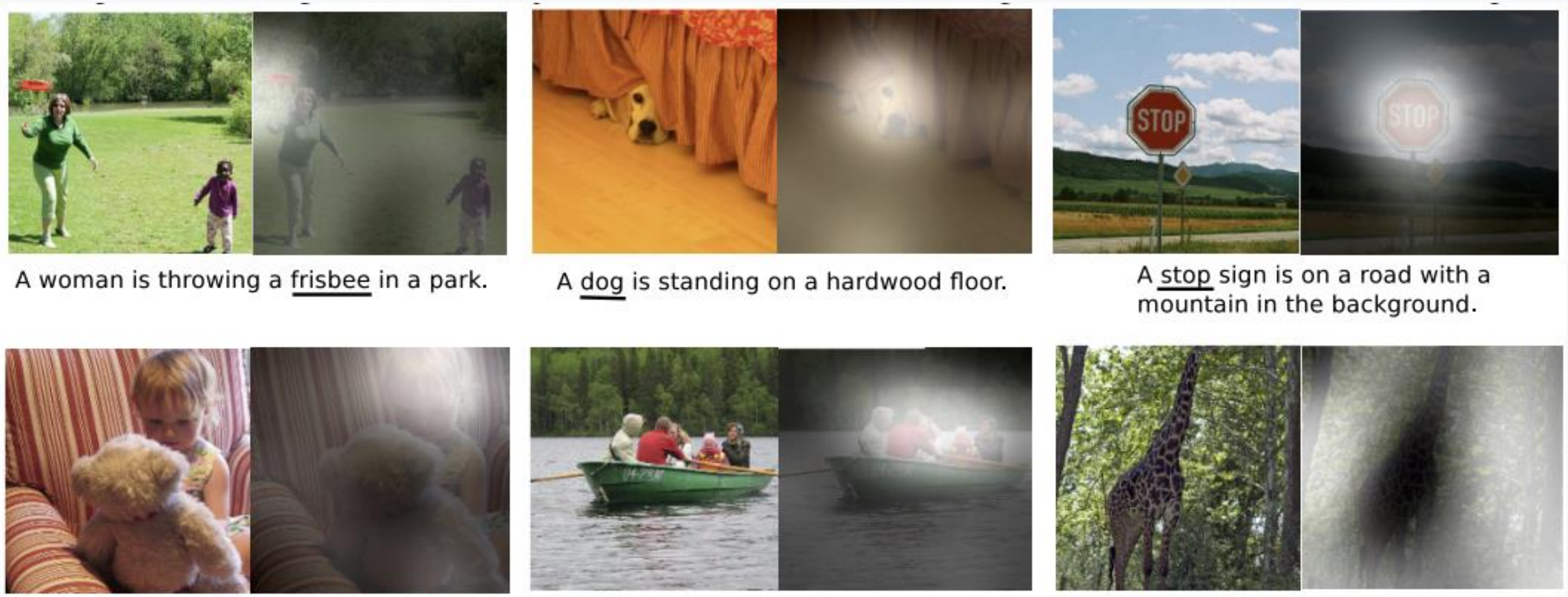


◆ Related Work and Problem Context

❖ Recent work

✓ Attention mechanism philosophy

- Interact with sound context and visual representation across spatial axes



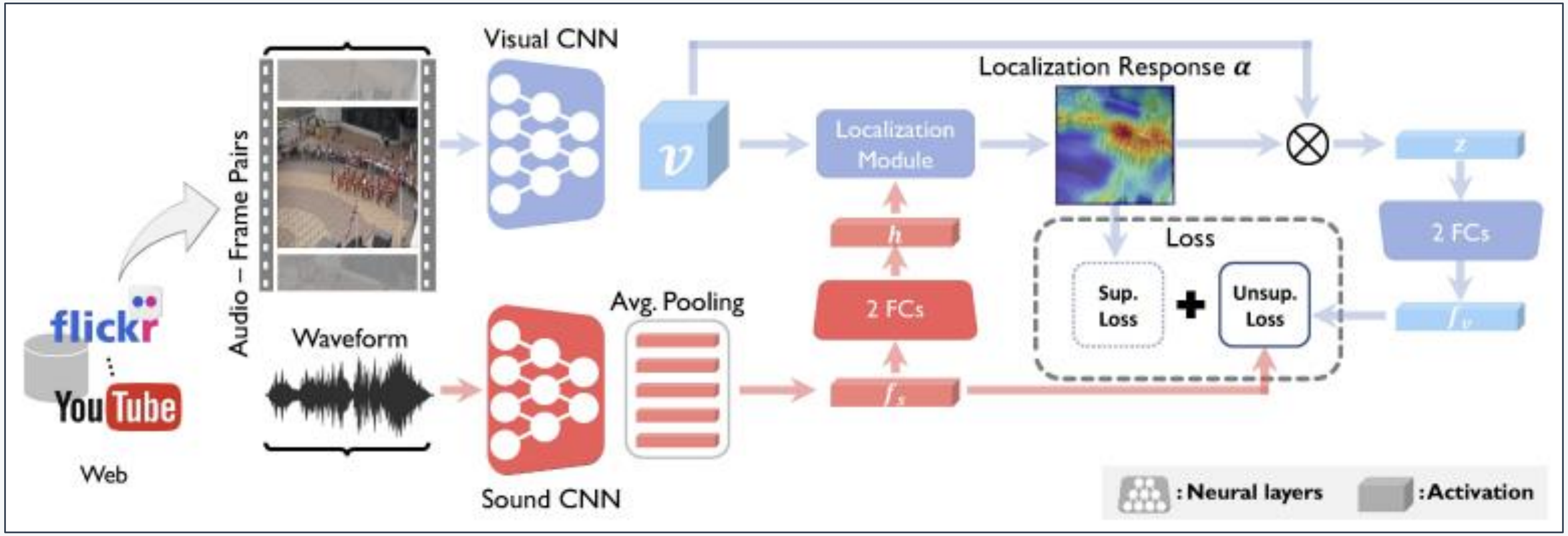
A woman is throwing a frisbee in a park.
A dog is standing on a hardwood floor.
A stop sign is on a road with a mountain in the background.
A little girl sitting on a bed with a teddy bear.
A group of people sitting on a boat in the water.
A giraffe standing in a forest with trees in the background.

Ref) Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." *International conference on machine learning*. PMLR, 2015.

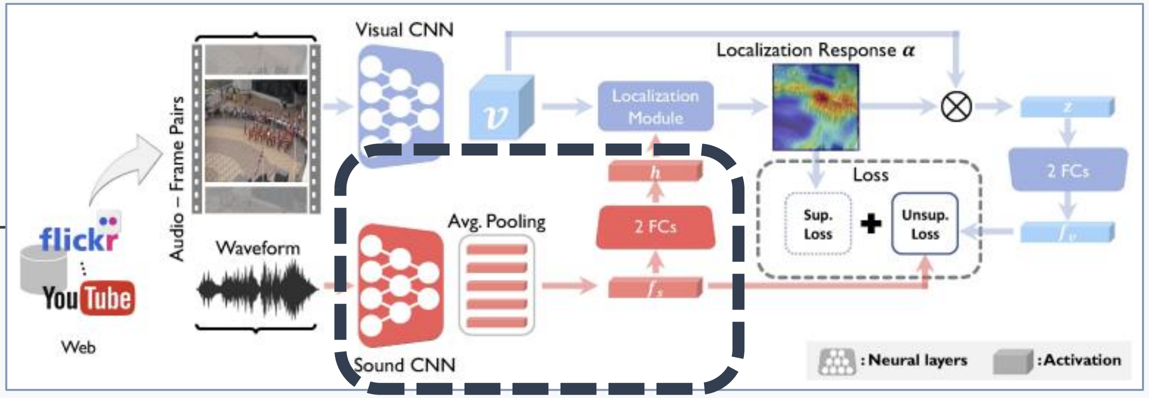
◆ Proposed Algorithm

❖ Vision based sound localization within the unsupervised learning framework

> Two-stream network architecture



- ✓ Sound network
- ✓ Visual network
- ✓ Attention model



◆ Proposed Algorithm

❖ Sound Network

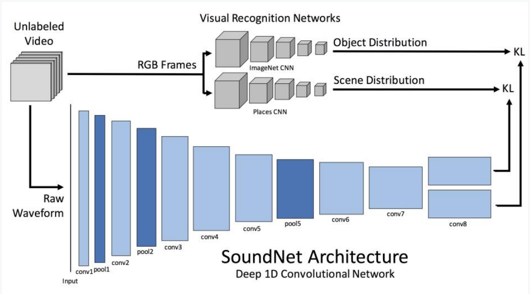
- Capture the concept of sound
- Convolutional module (**conv**), rectified linear unit (**ReLu**), pooling (**pool**) -> stack 10 layers
- 1-D deep convolutional architecture : invariant to input length
  - Use average pooling over sliding windows

Input

raw waveform

conv1 ~ conv8

SoundNet



+ 1000filters + Avg. Pooling
(sliding window 20sec)

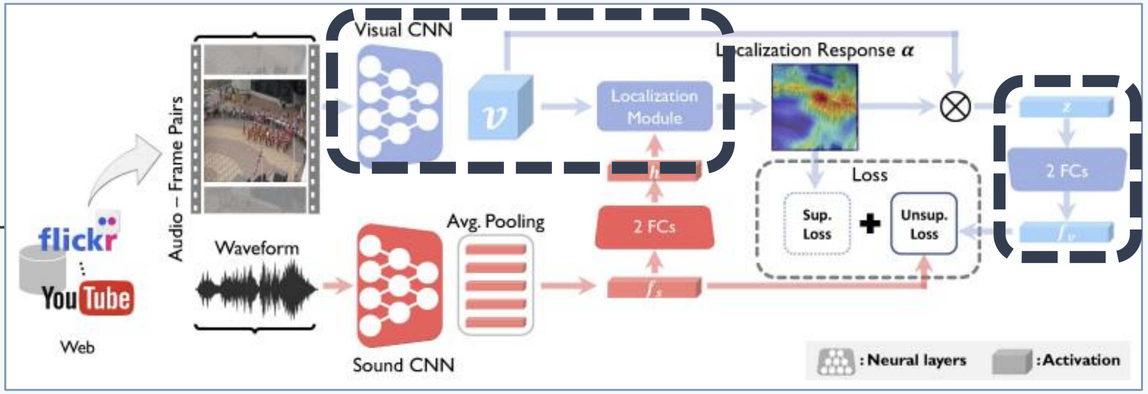
f_s
(1000-D)
sound representation

9, 10TH

ReLu-Fc layers

Output

h
(512-D)
sound embedding



◆ Proposed Algorithm

❖ Visual Network

- Image feature extractor + localization module
- Activation $\mathcal{V} \in \mathbb{R}^{H' \times W' \times D}$ ($H' = \left[\frac{H}{16}\right], W' = \left[\frac{W}{16}\right], D = 512$)
- Localization module : reveal sound source location information in the grid

Input

Color video frames (H x W)

~ conv5_3

VGG-16

$\mathcal{V}$ (512-D) activation

Localization module

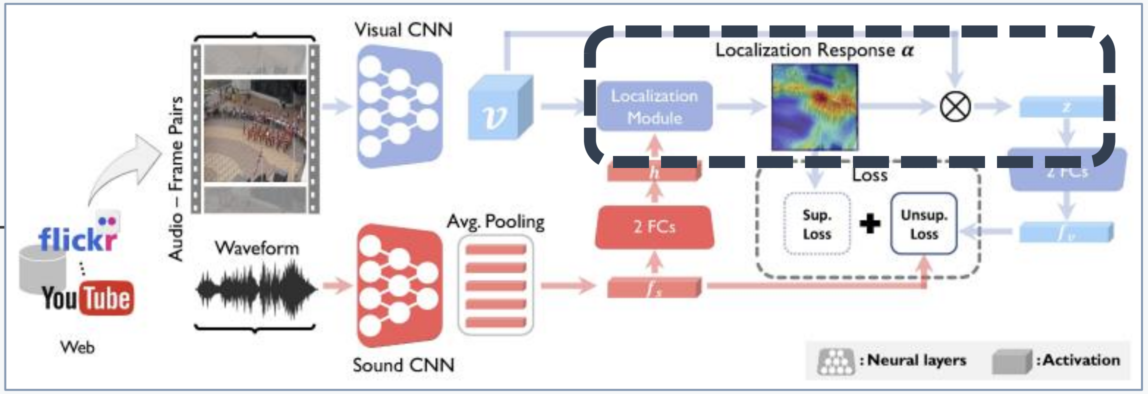
$\boldsymbol{h}$ (Sound embedding)

Confidence map of sound source + $\mathbf{z}$ (visual representation)

2 × ReLu-Fc

$\boldsymbol{f}_v$ (visual embedding)

Output



◆ Proposed Algorithm

❖ Localization Network

- Extracted visual and sound concepts -> localization networks

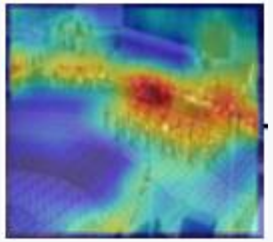
  ⇒ **sound source location** (a soft confidence score map)
- Modeled based on the **attention mechanism**
- Reshape $[v_1; \cdots; v_M] \in \mathbb{R}^{M \times D}$ $(\mathrm{M} = H'W')$
- **Attention α_i** : probability that the grid i - right location related to sound context (i ∈ {1,· · ·,M})

Attention α_i

$$\alpha_i = \frac{\exp(a_i)}{\sum_j \exp(a_j)}, \quad \text{where } a_i = g_{\mathtt{att}}(\mathbf{v}_i, \mathbf{h})$$

[Mechanism 1] $\quad g_{\cos}(\mathbf{v}_i, \mathbf{h}) = \bar{\mathbf{v}}_i^\top \bar{\mathbf{h}},$

[Mechanism 2] $\quad g_{\mathrm{ReLu}}(\mathbf{v}_i, \mathbf{h}) = \max(\bar{\mathbf{v}}_i^\top \bar{\mathbf{h}}, 0)$



map

Visual feature Z

$$\mathbf{z} = \mathbb{E}_{p(i|\mathbf{h})}[\hat{z}] = \sum_{i=1}^{M} \alpha_i \mathbf{v}_i$$

> The local visual feature at the sound source location

◆ Localizing Sound Source via Listening

> Video

> Video **frame**

prediction

> Audio signals - **sound**

prediction

◆ Localizing Sound Source via Listening

❖ Unsupervised learning

Features $\boldsymbol{f_v}$(video frame) and $\boldsymbol{f_s}$(sound wave)

✓ Corresponding pairs (positive) – close to each other -> extracted from same video

✓ Non-corresponding pairs (negative) – far from each other -> extracted from another random video

$\Rightarrow$ Triplet loss

> A triplet network

$$T(\underbrace{\mathbf{f}_v}_{\text{query}}, \underbrace{\mathbf{f}_s^-}_{\text{Negative sample}}, \underbrace{\mathbf{f}_s^+}_{\text{Positive sample}}) = [\|\mathbf{f}_v - \mathbf{f}_s^+\|_2, \|\mathbf{f}_v - \mathbf{f}_s^-\|_2] = \underbrace{[d_+, d_-]}_{\text{Two distance terms}}$$

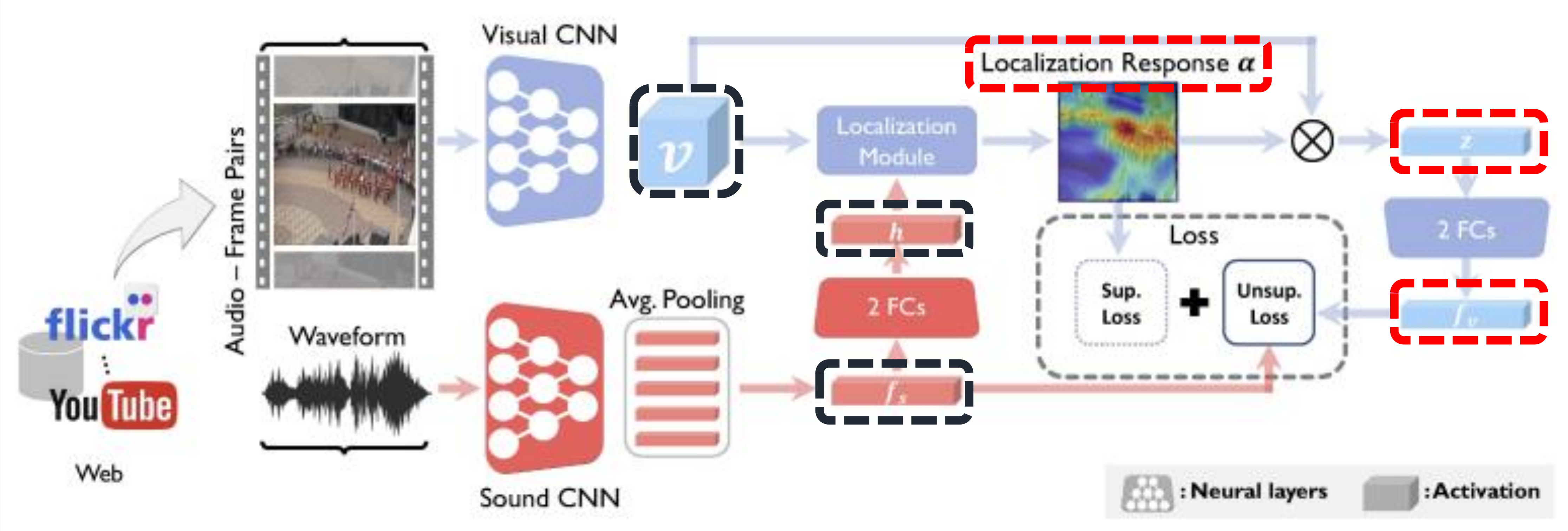
Two distance terms
$d_+ < d_-$

> Unsupervised loss function

$$\mathcal{L}_U(D_+, D_-) = \|D_+\|_2^2 + \|1 - D_-\|_2^2$$

$$D_\pm = \frac{\exp(d_\pm)}{\exp(d_+) + \exp(d_-)}$$

◆ Localizing Sound Source via Listening

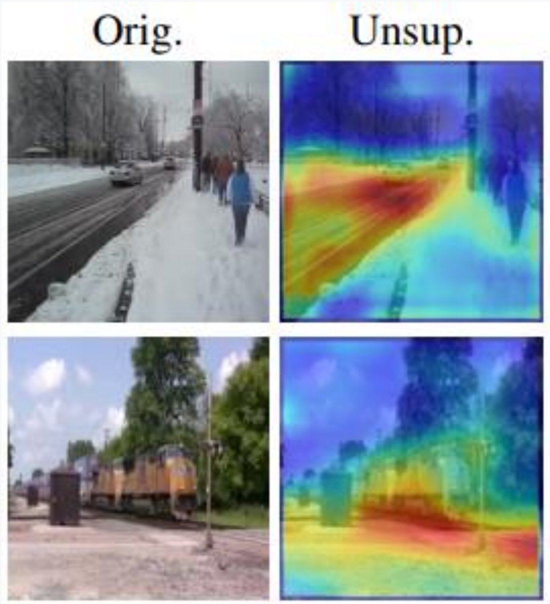


> A cycle loop

◆ Localizing Sound Source via Listening

❖ Unsupervised learning - issue

✓ The pigeon superstition issue



> Railways rather than train

⇒ False conclusion

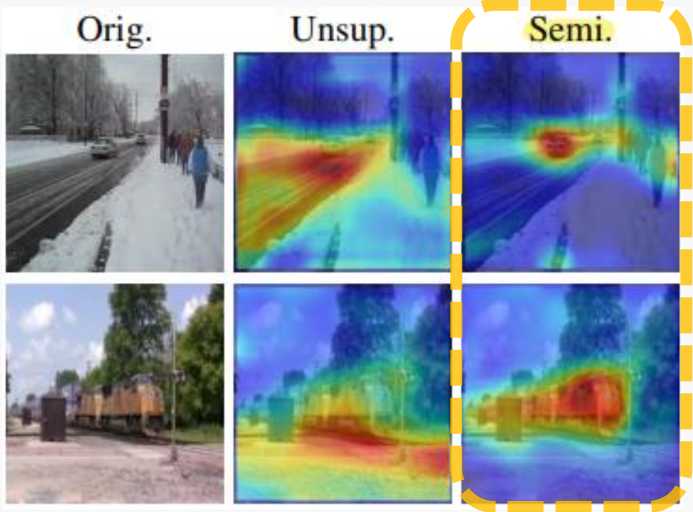
- Road > Car
- Difficult without supervisory feedbacks

⇒ **Biasing** toward a certain semantically unrelated output

◆ Localizing Sound Source via Listening

❖ Semi-supervised learning

- A small amount of prior knowledge – induce better inductive bias
- Add a supervised loss



> Semi-supervised loss function

Ground-truth attention map

$$\mathcal{L}(\mathbf{f}_v, \mathbf{f}_s^+, \mathbf{f}_s^-, \boldsymbol{\alpha}, \boldsymbol{\alpha}_{\mathrm{GT}}) = \mathcal{L}_U(\mathbf{f}_v, \mathbf{f}_s^+, \mathbf{f}_s^-) + \lambda(\boldsymbol{\alpha}_{\mathrm{GT}}) \cdot \mathcal{L}_S(\boldsymbol{\alpha}, \boldsymbol{\alpha}_{\mathrm{GT}})$$

Supervised loss

> Cross entropy loss

$$\mathcal{L}_S(\boldsymbol{\alpha}, \boldsymbol{\alpha}_{\mathrm{GT}}) = -\sum_i \alpha_{\mathrm{GT},i} \log(\alpha_i)$$

Unsupervised loss

Function for controlling the data supervision type

⇒ ground-truth existence

$\lambda(X) = 0 \;\; if \; X \in \emptyset,$

$\lambda(X) = 1 \;\; otherwise$

◆ Experimental Results

❖ Dataset

✓ Unlabeled **Flickr-SoundNet dataset**

- more than 2 million unconstrained sound & image pairs
- Random subset of 144k pairs - train

✓ **Annotated in image coordinates**

- Training supervision models
- 5k frames + sound → 3 subjects

1. Listen 20 secs
2. Draw bounding box
3. Tag the bounding box as object or ambient

- 2786 pairs => 250 – Testing, 2236 – Training

> Annotated

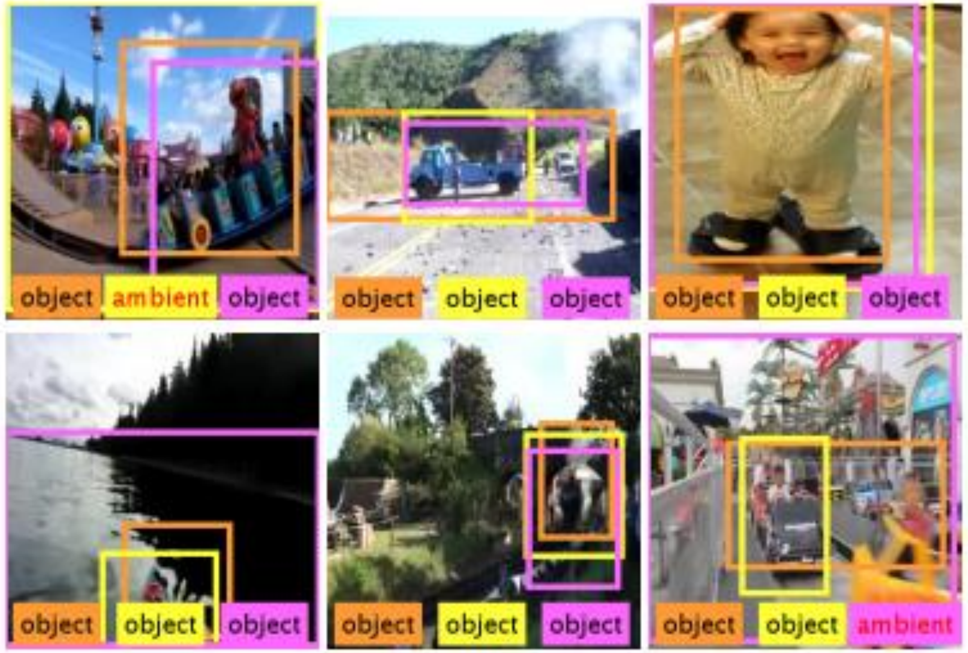
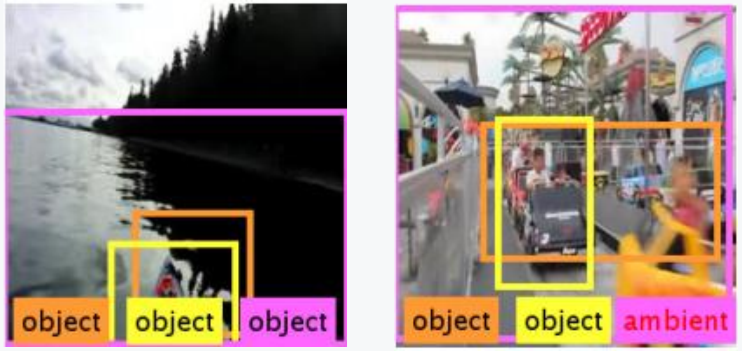


Figure 4. **Sound Source Localization Dataset.** Human annotators annotated the location of the sound source and the type of the source (object vs. non-object/ambient). This dataset is used for testing how well our network learned the sound localization and also for providing a supervision to unified architecture.

◆ Experimental Results – results and analysis

❖ Evaluation metrics

✓ **3 annotations** from 3 respective subjects



> ambiguous

✓ **Weighted score map**

1. Bounding box annotation → binary maps $\{b_j\}_{j=1}^{N}$
2. Extract a representative score map $\mathbf{g}$

$$\mathbf{g} = \min\left(\sum_{j=1}^{N} \frac{\mathbf{b}_j}{\#\text{consensus}}, 1\right)$$

$N = 3$

Min. opinion to agreement = 2

⇒ Positives ≥ #consensus → $\mathbf{g} = 1$

⇒ **Consensus intersection over union (cIoU)**

$$\text{cIoU}(\tau) = \frac{\sum_{i \in \mathcal{A}(\tau)} g_i}{\sum_i g_i + \sum_{i \in \mathcal{A}(\tau) - \mathcal{G}} 1}$$

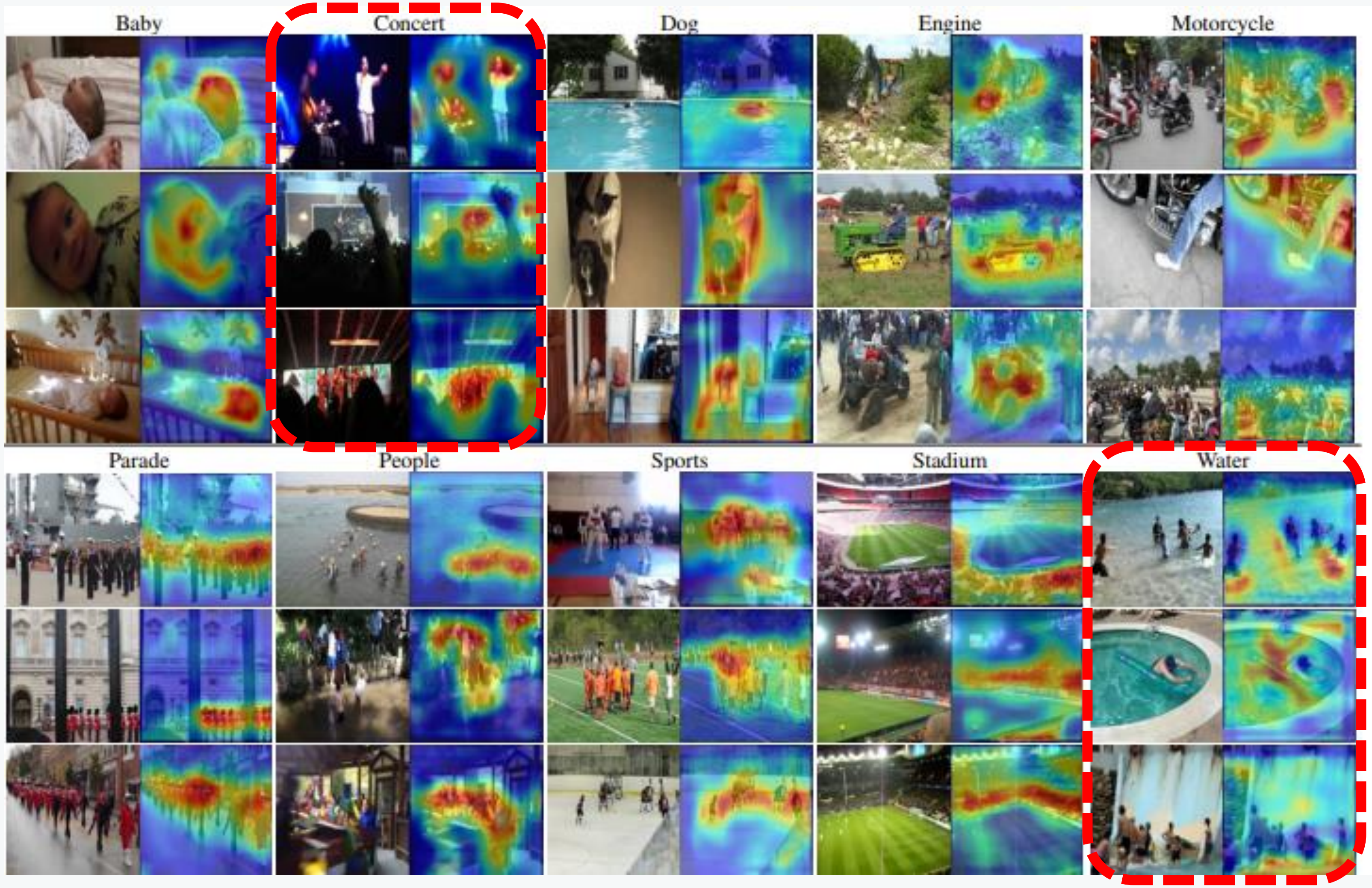
$\mathcal{A}(\tau) = \{i | \alpha_i > \tau\}$

$\mathcal{G} = \{i | g_i > 0\}$

◆ Experimental Results – results and analysis

❖ Qualitative Analysis

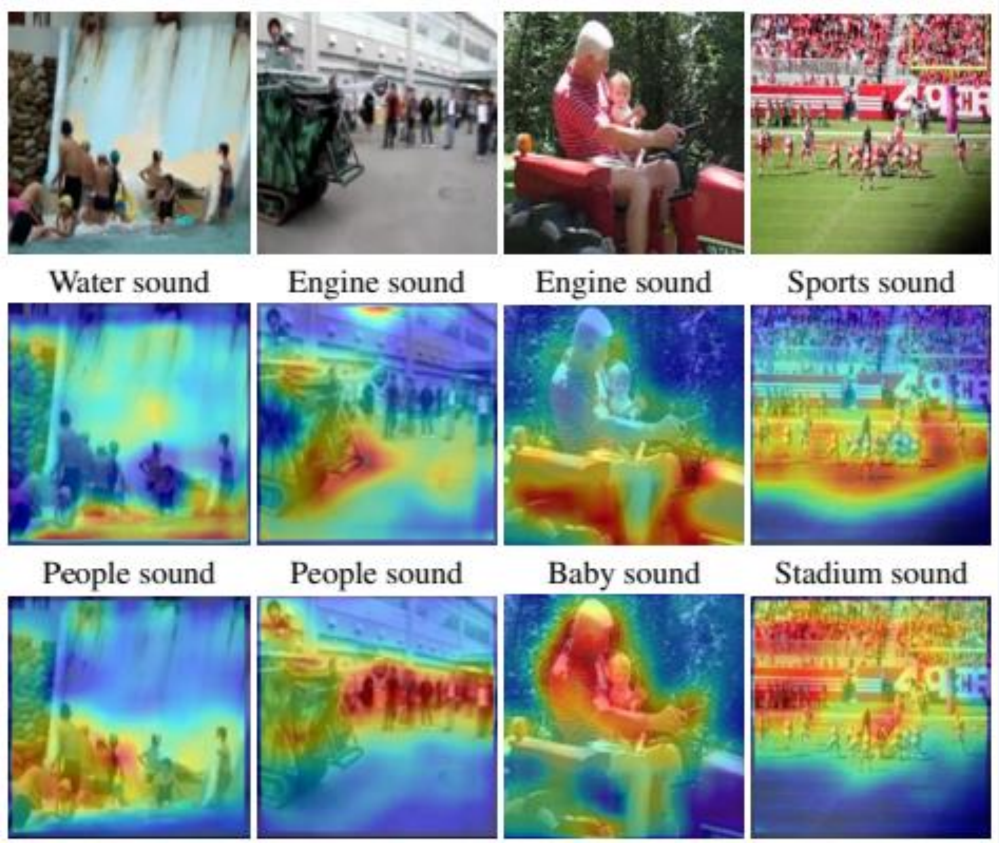
⇒ Visualize Localization response α

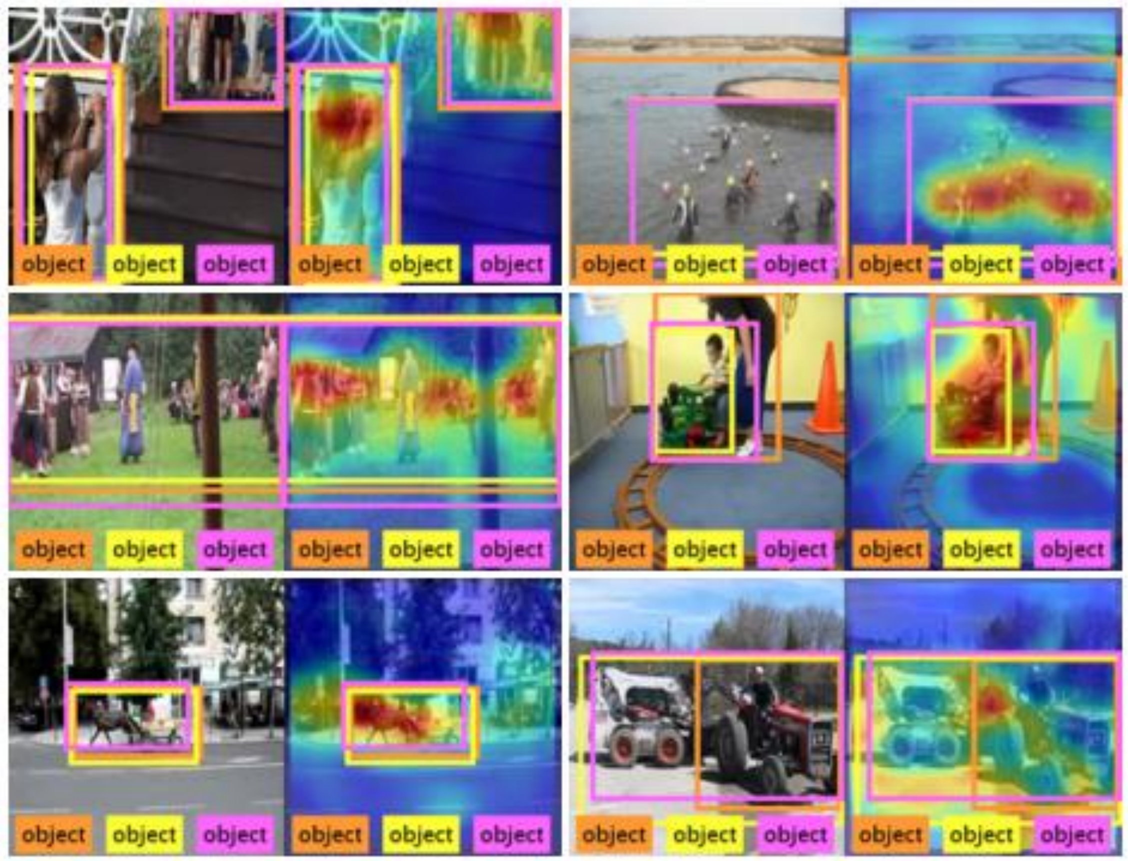


> **Unsupervised** Network

◆ Experimental Results – results and analysis

❖ Qualitative Analysis



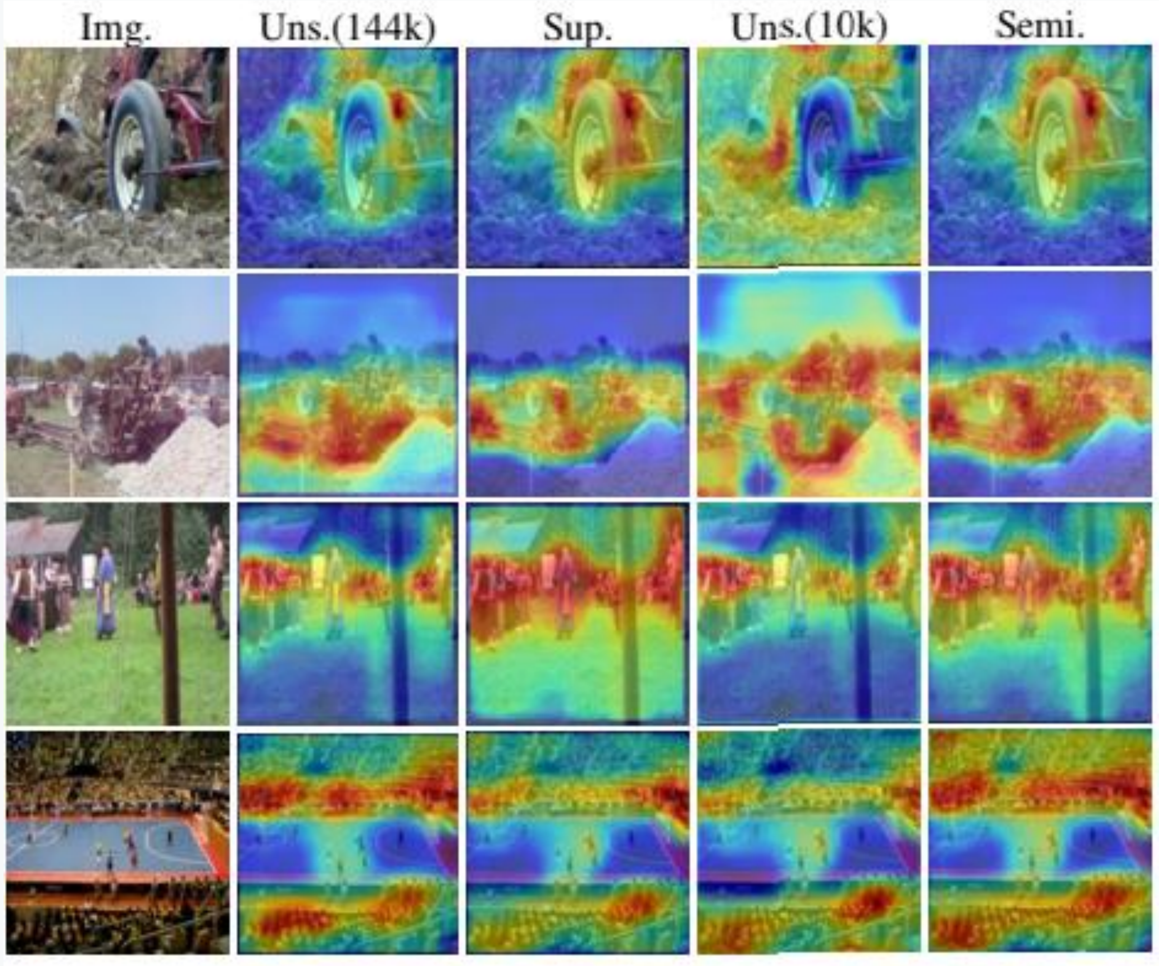


\> **Interactive** sound source localization

\> Our Network **vs Human annotation**

◆ Experimental Results – results and analysis

❖ Qualitative Analysis



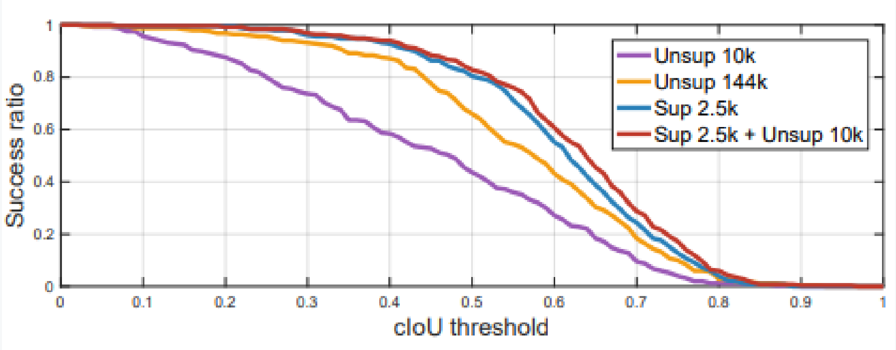
> Different Learning Methods

◆ Experimental Results – results and analysis

❖ Quantitative Analysis

✓ Table 1

	softmax		ReLU+softmax	
	cIoU	AUC	cIoU	AUC
Unsup. 10k	43.6	44.9	–	–
Unsup. 144k	66.0	55.8	52.4	51.2
Sup. 2.5k	80.4	60.3	82.0	60.7
Sup. 2.5k + Unsup. 10k	82.8	62.0	84.0	61.9
Random	cIoU		AUC	
	0.12 ± 0.2		32.3 ± 0.1	



> Performance evaluation with different learning schemes

✓ Table 2

	softmax		ReLU+softmax	
	cIoU	AUC	cIoU	AUC
Unsup. 10k	43.6	44.9	–	–
Unsup. 144k	66.0	55.8	52.4	51.2
Sup. 0.5k + Unsup. 10k	78.0	60.5	79.2	60.3
Sup. 1.0k + Unsup. 10k	82.4	61.1	82.4	61.1
Sup. 1.5k + Unsup. 10k	82.0	61.3	82.8	61.8
Sup. 2.0k + Unsup. 10k	82.0	61.5	82.4	61.4
Sup. 2.5k + Unsup. 10k	82.8	62.0	84.0	61.9

> Semi-supervised learning with different number of samples

✓ Table 3

Subject	Unsup. 144k		Sup.		Semi-sup.	
	IoU	AUC	IoU	AUC	IoU	AUC
Subj. 1	58.4	52.2	70.8	55.6	74.8	57.1
Subj. 2	58.4	52.4	72.0	55.6	73.6	57.2
Subj. 3	63.6	52.6	74.8	55.6	77.2	57.3
Avg.	60.1	52.4	72.5	55.6	75.2	57.2

> Performance measure against individual subjects

◆ Discussion and Conclusion

✓ **Learning based sound source localization in visual scenes** – build its new benchmark dataset

- The model plausibly works / can often get to false conclusion without prior knowledge.
- Leveraging small amount of human knowledge can correct to capture semantically meaningful relationships

⇒ The task is not fully learnable problem only with unsupervised data, but it can be fixed by providing even small amount of supervision.

> Pigeons issue

- To deduce the way of machine learning about sound source localization in visual scenes.
- Sound based representation learning : at least small amount of supervision should be incorporated
- Open many potential directions for future research
  - Multi-modal retrieval
  - Sound based saliency or representation learning and its applications.

Ref) R. Arandjelovic and A. Zisserman. Look, listen and learn. In ´ IEEE International Conference on Computer Vision, 2017