졸업프로젝트 발표

2019101230 국제학과 신희연

Contents

1) **Background**: What is Audio-Visual Segmentation(AVS)?

2) **Related Conference Paper**: “Audio-Visual Segmentation” (2022 ECCV)

3) **Summary of Our Research**

   → Background(Purpose)

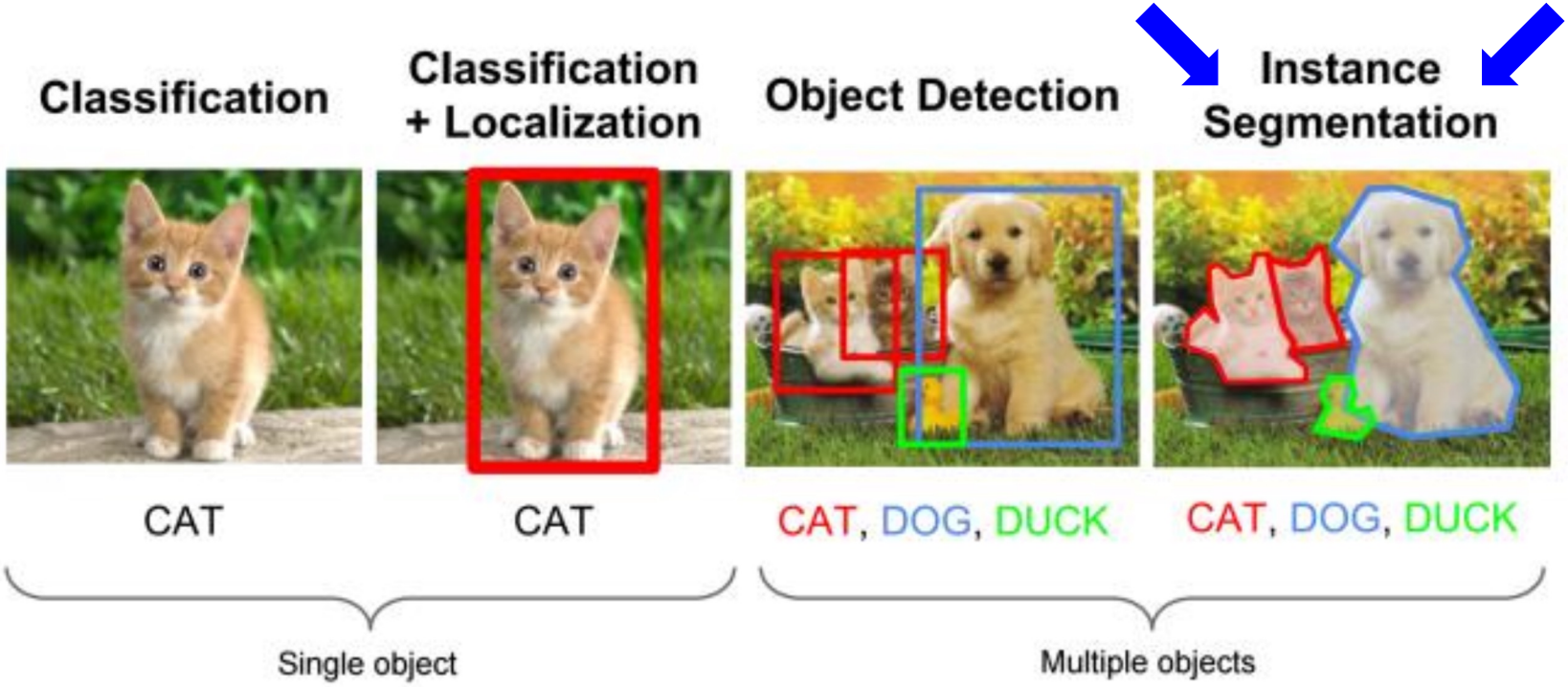
   → Methodology

   → Experiment

   → Conclusion

4) References

What is Audio-Visual Segmentation(AVS)?



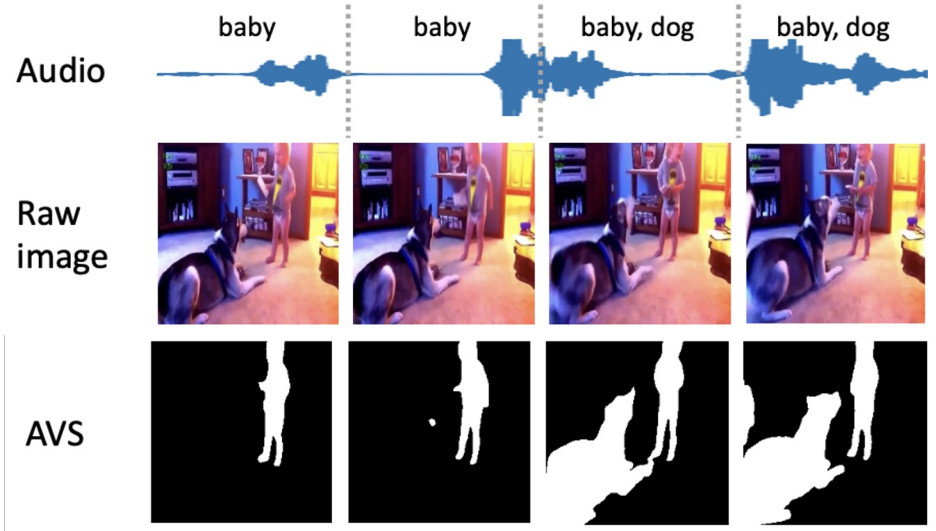
What is Audio-Visual Segmentation(AVS)?

- **Image Segmentation** applied to the Multi-modal (Audio-Visual) field
- First introduced at **ECCV 2022 : 'Audio-Visual Segmentation'**

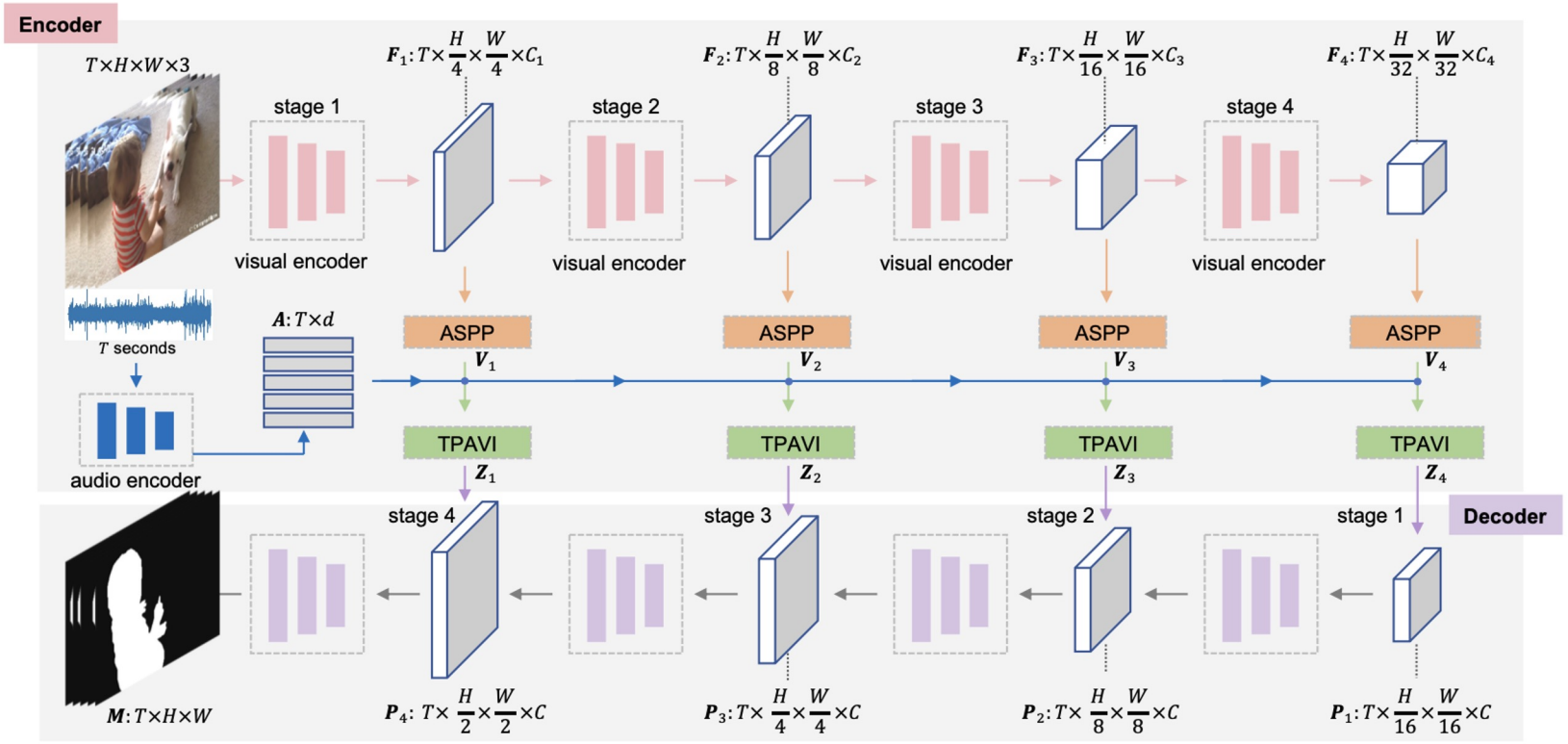
original

Semantic Seg.

Instance Seg.



Review: Audio-Visual Segmentation



Review: Audio-Visual Segmentation

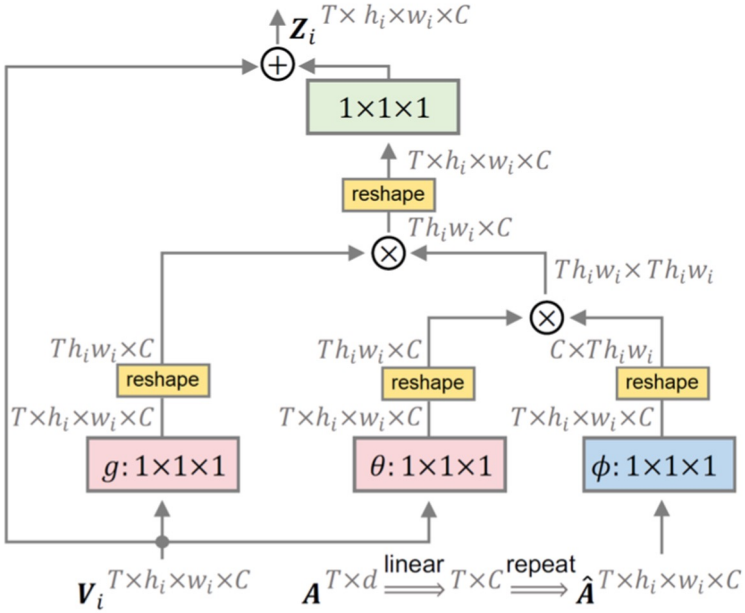


Fig. 5. The TPAVI module

$$\boldsymbol{Z}_i = \boldsymbol{V}_i + \mu(\alpha_i\ g(\boldsymbol{V}_i)),\ \text{where}\ \alpha_i = \frac{\theta(\boldsymbol{V}_i)\ \phi(\hat{\boldsymbol{A}})^\top}{N}$$

V_i : visual feature

N : $T\times h_i\times w_i$, normalized vector

α_i : audio-visual similarity

$\hat{A}$: processed audio feature

Z_i : final output of TPAVI module

Review: Audio-Visual Segmentation

Objective function	MS3 ($\mathcal{M}_{\mathcal{J}}$)		MS3 ($\mathcal{M}_{\mathcal{F}}$)	
	ResNet50	PVT-v2	ResNet50	PVT-v2
$\mathcal{L}_{\text{BCE}}$	.466	.531	.558	.626
$\mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{AVM-VV}}$	.467	.538	.577	.644
$\mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{AVM-AV}}$	**.479**	**.540**	**.578**	**.645**

$$\mathcal{L} = \text{BCE}(\boldsymbol{M}, \boldsymbol{Y}) + \lambda \mathcal{L}_{\text{AVM}}(\boldsymbol{M}, \boldsymbol{Z}, \boldsymbol{A}),$$

$$\mathcal{L}_{\text{AVM}} = \sum_{i=1}^{n} (\text{KL}(avg\ (\boldsymbol{M}_i \odot \boldsymbol{Z}_i), \boldsymbol{A}_i)),$$

M : prediction (final output of decoder)

Z : output of TPAVI module

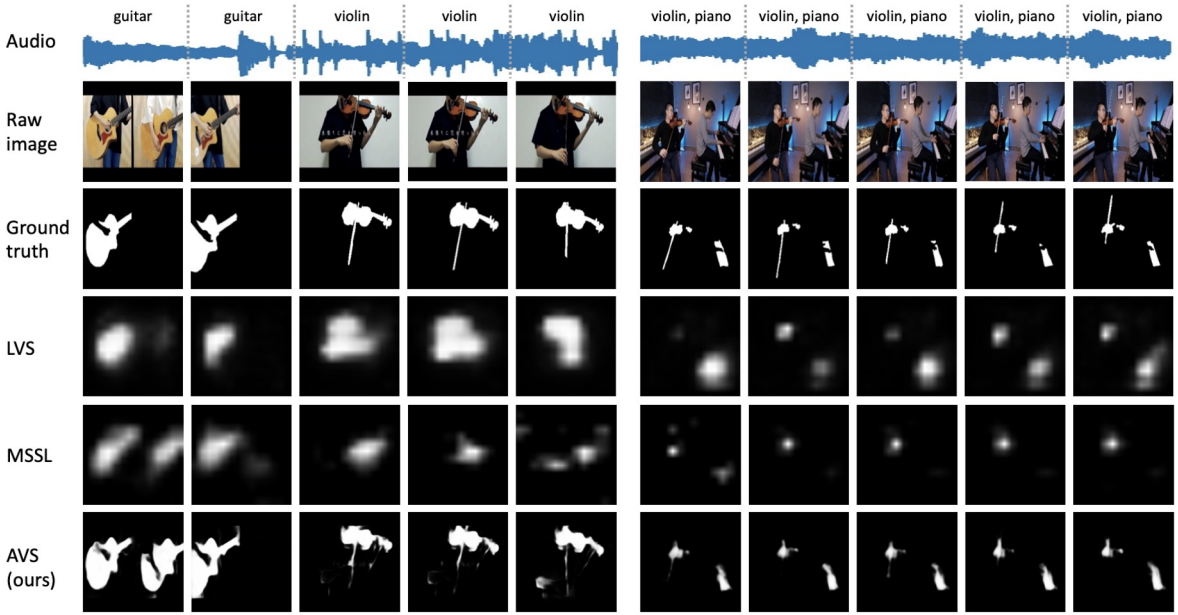
A : audio feature

Y : pixel-wise label

$\odot$: matrix multiplication

Review: Audio-Visual Segmentation

Metric	Setting	SSL		VOS		SOD		AVS (ours)	
		LVS[5]	MSSL[30]	3DC[27]	SST[10]	iGAN[28]	LGVT[49]	ResNet50	PVT-v2
$\mathcal{M}_{\mathcal{J}}$	S4	.379	.449	.571	.663	.616	.749	.728	**.787**
	MS3	.295	.261	.369	.426	.429	.407	.479	**.540**
$\mathcal{M}_{\mathcal{F}}$	S4	.510	.663	.759	.801	.778	.873	.848	**.879**
	MS3	.330	.363	.503	.572	.544	.593	.578	**.645**



Mean Metric Values

MJ: computes the intersection-over-union(IoU) of the predicted segmentation and the ground truth mask

MF: considers both the precision and recall

$$\frac{(1+\beta^2)\times \text{precision}\times \text{recall}}{\beta^2\times \text{precision}+\text{recall}}$$

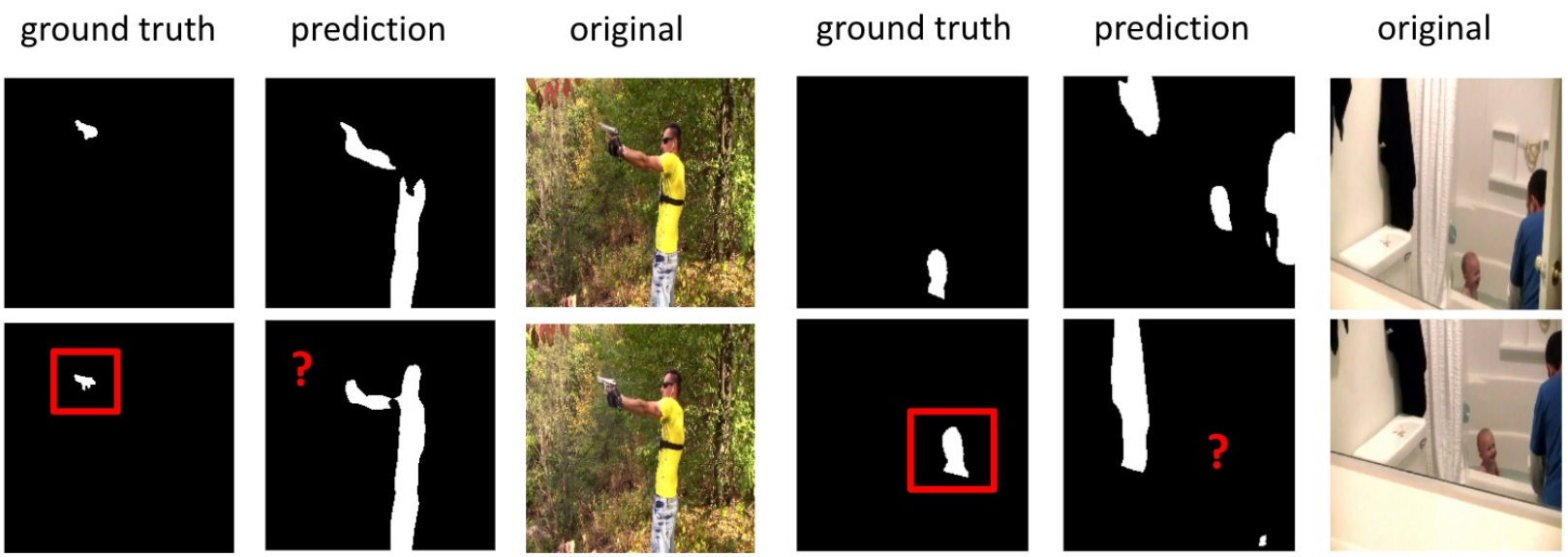
where β^2 is set to 0.3

Summary of Our Research: Background

Purpose

그런데, 실제로 AVSModel이 단순히 이미지 내 객체를 탐지하는 것이 아니라,
구체적으로 **소리가 나는 객체**를 탐지하고 있는가?

Summary of Our Research: Background



Summary of Our Research: Background

Ideas

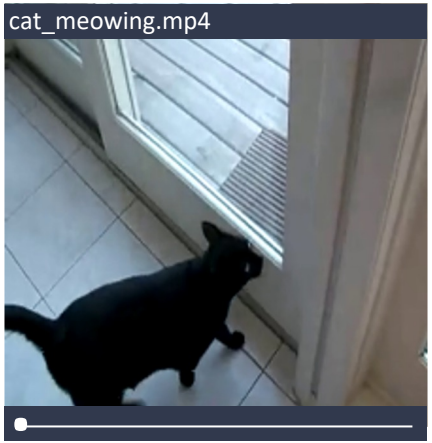
소리가 나는 객체를 원본 이미지에서 Crop하여 소리와 엮는 방법

1. Object Classification with Cropped Images → Classification

   → Loss 를 추가하였습니다.

2. Audio features를 Cropped Images와 함께 고려하여 Loss를 추가하였습니다.

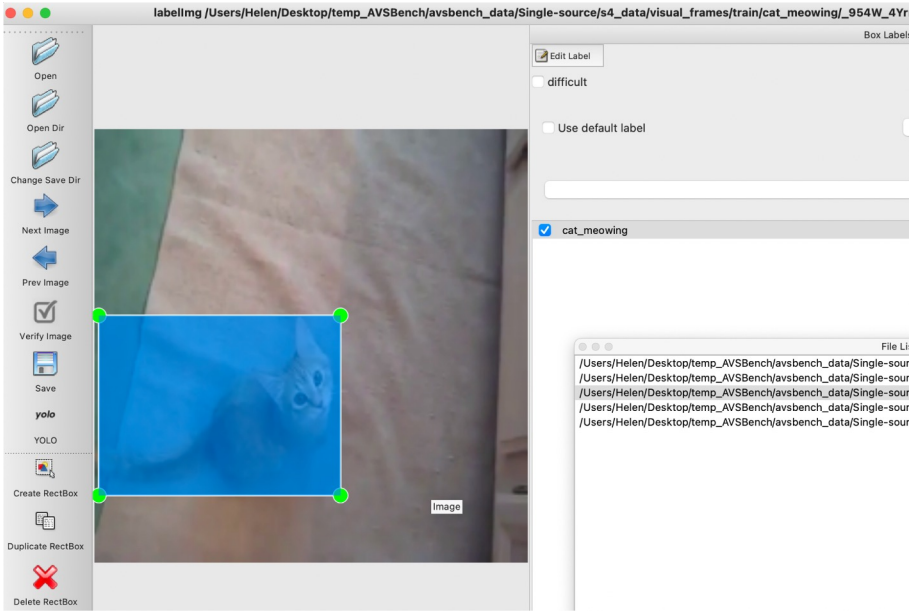
Summary of Our Research: Methodology



5초 길이의 비디오를 5개의
visual frames 로 분할

Summary of Our Research: Methodology

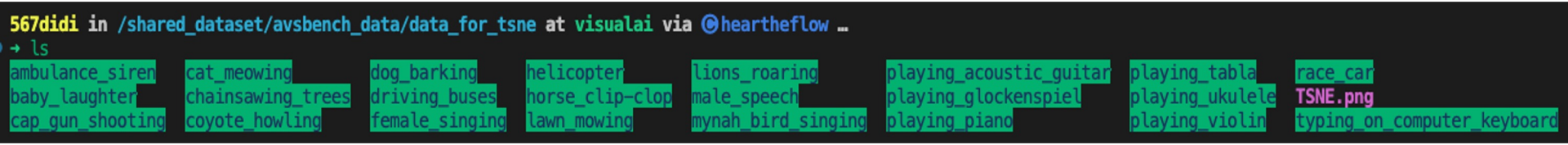
Bounding Box



Cropping & Resizing

Summary of Our Research: Methodology

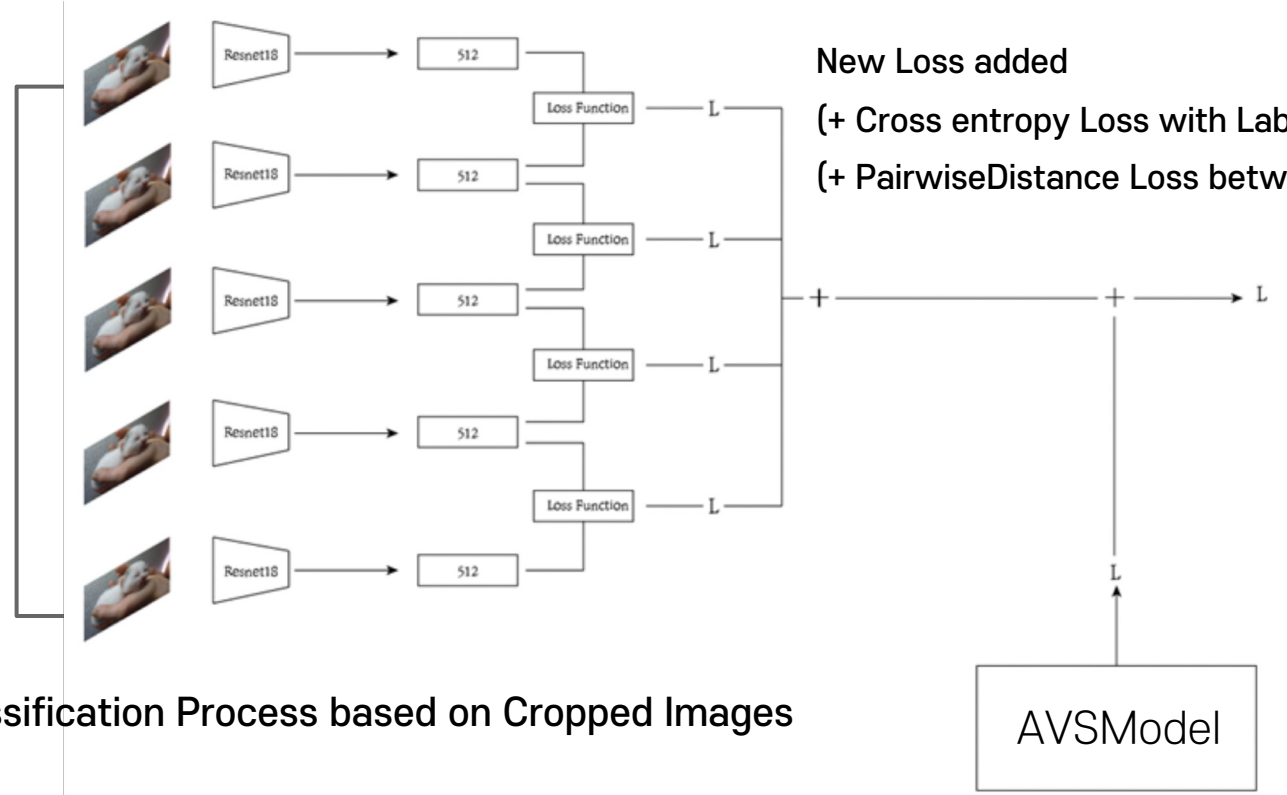
Single Source Dataset 총 17,260장(23개 Class) Labeling 완료



Summary of Our Research: Methodology

먼저, 5개의 Cropped Image를 Input으로 ResNet50 → Classification Loss 추가하였습니다.

Contrastive Learning



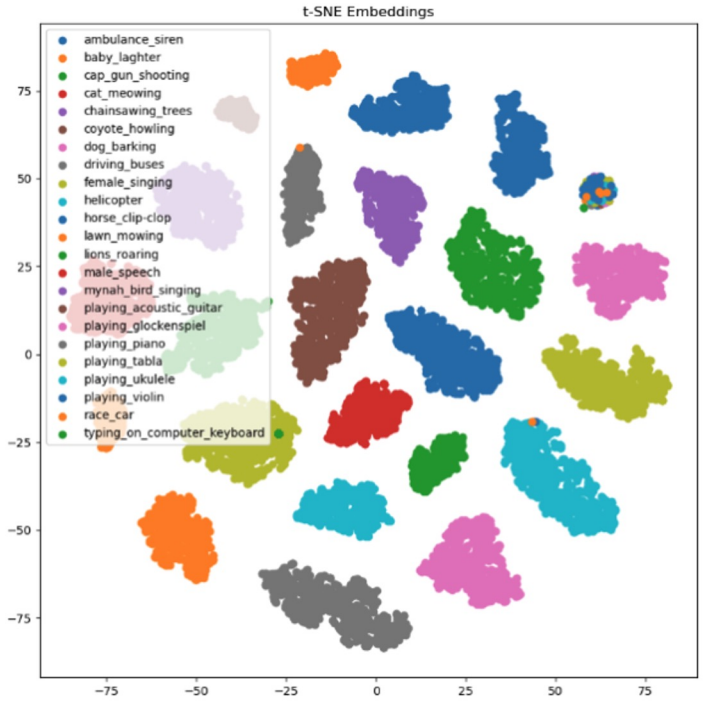
New Loss added

(+ Cross entropy Loss with Label)

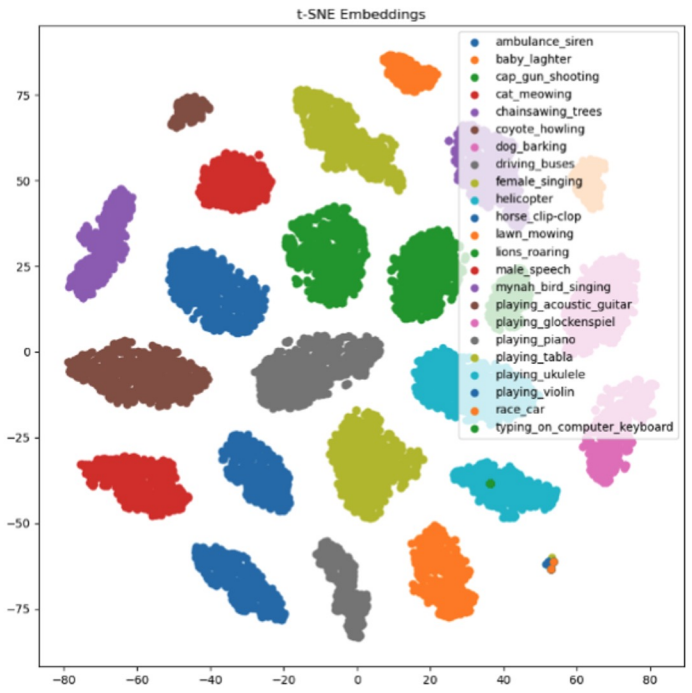
(+ PairwiseDistance Loss between Images)

Classification Process based on Cropped Images

Summary of Our Research: TSNE Visualization



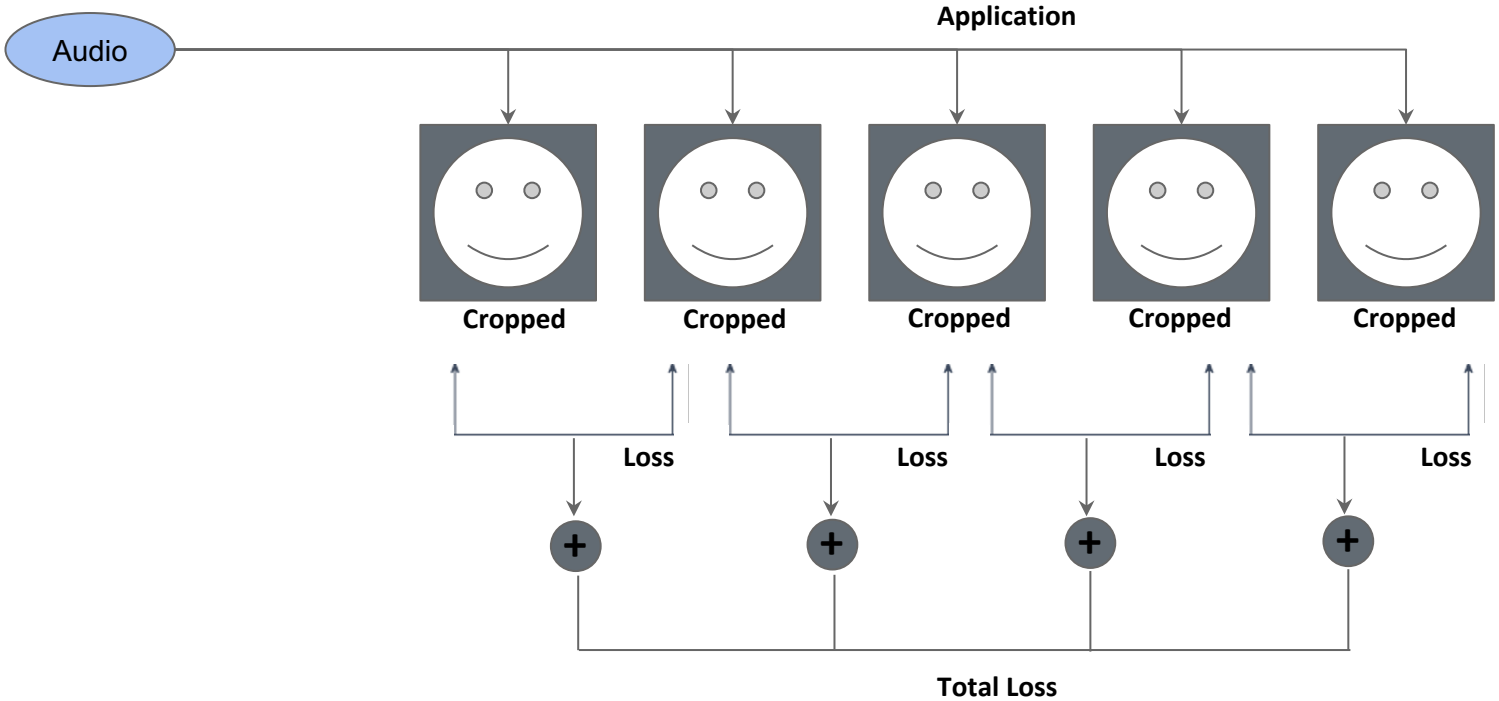
Epoch: 10



Epoch: 100

Summary of Our Research: 2nd Experiment

이후, 오디오 Feature도 Loss와 함께 고려되기 위해 1) 5개의 오디오를 평균낸 후
2) 각각의 Cropped Image와 **코사인 유사도**를 구하여 새 Loss에 추가한 방법입니다.



Summary of Our Research: Experiment

기존 논문과 오디오 및 이미지 Feature를 고려한 Loss를 추가한 결과를 비교해보았을 때,
기존 논문 성능에 비해 성능이 향상된 것을 볼 수 있습니다.

ResNet 50	논문 기준 TPAVI	**Proposed Method** (Audio, Visual에 관한 Loss를 추가한 방법)
Epoch	15	15
Train: Best Miou	0.682 at Epoch 8	0.7196 at Epoch 9
Test: Best Miou	0.679	0.726
Test: F_score	0.79	0.8449

Summary of Our Research: Conclusion

즉,

1) Cropped Image의 Classification 학습 Loss와

2) 5개의 오디오를 평균 낸 후, 각 5개의 Cropped Image와

코사인 유사도를 구하여 얻은 Loss를 더하였을 때

최종 Audio-Visual Segmentation 성능이 향상된 것을 알 수 있었습니다.

Summary of Our Research: Conclusion

감사합니다

References

[1] labelImg, https://github.com/heartexlabs/labelImg

[2] T. Baltrušaitis, C. Ahuja and L. -P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423-443

[3] Senocak, A., Oh, T. H., Kim, J., Yang, M. H., & Kweon, I. S. (2018). Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*

[4] Valverde, F. R., Hurtado, J. V., & Valada, A. (2021). There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

[5] Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., ... & Zhong, Y. (2022, October). Audio–Visual Segmentation. In *Computer Vision–ECCV 2022: 17th European Conference*

[6] Ziegler, A., & Asano, Y. M. (2022). Self-supervised learning of object parts for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*