Master thesis on Sound and Music Computing

Universitat Pompeu Fabra

Multimodal Urban Scene Understanding

Rajsuryan Singh

Supervisors:

Pablo Zinemanas

Magdalena Fuentes

August 2022

Table of Contents

Abstract

Early computational approaches for sound source localization, originating in robotics, were modeled after animal perception and utilized audiovisual synchrony and spatial information inferred from multichannel audio. More recent deep learning-based methods focus on learning semantic audiovisual representations in a self-supervised manner and using them for localizing sounding objects. A majority of these approaches by design exclude information that comes from the temporal context that a video provides. While that is not a hurdle for widely used benchmark datasets because of the bias towards having large single objects in the middle of the image, the methods fall short on more challenging scenarios like urban traffic videos. This thesis aims to explore methods to introduce temporal context into the state-of-the-art methods for sound source localization in urban scenes. Optical flow is used as a means to encode motion information. An analysis of the strengths and weaknesses of our methods helps us better understand the problem of visual sound source localization and sheds new light on the characteristics of our dataset.

Keywords: visual sound source localization; urbansas; self-supervised learning; optical flow

Chapter 1

Introduction

Vision and audition are complementary sources of information, and effective integration of these senses offers undeniable survival advantages to an organism; the ability to localize sounds and connect them to visual objects enables a rich understanding of a dynamic environment. Sound source localization independent of visual inputs has been widely studied in robotics, generally using an array of microphones and classical signal processing techniques to estimate the location of a sound source [1, 2], somewhat mimicking animal perception. There has been a recent surge of interest in the problem following the release of datasets [3] and challenges [4] for sound event localization and detection (SELD). However, there is a continued and exclusive reliance on audio to localize and classify sound events which comes with certain limitations. Reverberation, low spatial resolution, interference, polyphony, and non-stationarity of sound sources have been shown to be severely detrimental to the performance of SELD systems [5].

Incorporating vision as a complementary modality offers a way to abate some of these limitations. Integrating audio and visual inputs allows, at least in principle, to attribute sounds to objects in a scene. It also opens up the possibility to leverage the rich body of work in computer vision to aid localization. Early attempts at modeling audio-visual perception exploited the synchrony between audio and visual events e.g. lip movements aligned to speech with probabilistic models like mutual information [6, 7], and canonical correlation analysis [8]. Following recent advances in deep learning, especially in computer vision, the field has pivoted to deep-neural-network-based methods which will be elaborated upon in the following sections.

A notable difference between the two approaches is the shift from using the temporal correlation between audio and video to the semantic similarity between them as the

primary source of information for localization. This has happened to the extent that most state-of-the-art methods completely disregard the temporal context available in videos [9-14]. These methods have focused on learning semantic auditory and visual representations in a self-supervised manner that enables sound source localization via the similarity between audio-visual embeddings. In layman's terms, if the audio contains the sound of a piano, the localization model would look for regions in the image that look like a piano and will attribute the sound to that region. In most cases, there is no explicit classification of sounds and images into discrete classes like a piano or a car. The audio and visual embeddings are optimized in a way that pushes similar classes close together while driving dissimilar ones farther apart in the representation space and the distance between the audio and visual embeddings in this space is used to measure the correspondence between them. This approach has been effective for the widely used benchmark datasets [9-14], however recent work by Ho-Hsiang, et al. has raised questions about the generalizability of these methods beyond these datasets [15]. They also point out the biases in these benchmarks and demonstrate that the methods developed on these datasets fail to generalize to urban scene understanding.

Automatic urban scene understanding is a growing area of research, with many potential applications in the industry, academia, and the public sector. The applications include assistive devices for the hearing-impaired, traffic monitoring, and autonomous driving. In addition to the practical applications, urban scenes provide a challenging scenario for visual sound source localization where state-of-the-art methods prove to be inadequate [15]. This motivates further research into the limitations of state-of-the-art methods as well as the datasets used for their development and that is what we set out to do in this thesis. The focus of this thesis is visual sound source localization in urban scenes and the main contribution is the incorporation of temporal information into sound source localization algorithms for urban scene understanding.

In this work, we develop and evaluate our methods on the Urban-Sound-and-Sight (Urbansas) dataset [16]. We use RC-Grad [15], which is the current state-of-the-art model for visual SSL for Urbansas, as our baseline. We propose the use of optical flow as a means to incorporate temporal information and we explore hard-coded as well as learning-based algorithms to combine it with RC-Grad. First, we use optical flow as a

heuristic to filter stationary objects from the predictions of RC-Grad and observe a significant improvement in localization performance, especially towards curbing false positives. We also analyze failure cases of the approach to get further insight into the robustness of optical flow and the factors that affect optical flow estimation. Further we add optical flow as a feature into the neural network in two ways - we add optical flow as an additional channel into the vision encoder, and train a separate optical flow encoder within the RC-Grad framework. A subsequent exploration of the strengths and weaknesses of our methods helps us better understand the limitations of the dataset and evaluation metrics we used and opens up lines of further inquiry.

The remaining chapters of this thesis are organized as follows. Chapter 2 starts with a general overview of self-supervised representation learning and then presents a review of the state-of-the-art in visual sound source localization. In Chapter 3, we justify the use of optical flow as a way to represent motion information in the context of urbansas and outline the methodology we propose to incorporate it into visual sound source localization algorithms. We present our results in Chapter 4 and we conclude with a discussion on the merits and limitations of our methods along with proposals for future work in Chapter 5.

Chapter 2

State of the Art

2.1. Representation learning

In the decade since Alexnet [17], the first of the innumerably many tours de force of deep learning, the field has burgeoned at an unprecedented rate. Deep learning systems have rivaled, and in some cases even surpassed, human-level performance in computer vision [17], audition [18], natural language processing[19], sensory prediction[20], game playing [21], and reasoning [22]. In a review published while deep learning was still in its infancy, Yoshua Bengio argues that the performance of machine learning methods relies heavily on the choice of data representation [23], and the unreasonable effectiveness of deep learning [24] comes from the ability to learn task-relevant representations from the data. While Bengio's review gives a rigorous and comprehensive account of the importance of representations, a much more accessible explanation for why representations are crucial for the success of an algorithm comes from David Marr [25]

> ..if one chooses the Arabic numeral representation, it is easy to discover whether a number is a power of 10 but difficult to discover whether it is a power of 2. If one chooses the binary representation, the situation is reversed. Thus, there is a trade-off; any particular representation makes certain information explicit at the expense of information that is pushed to the background and may be quite hard to discover. This issue is important because how information is represented can greatly affect how easy it is to do different things with it

In modern-day deep learning, the choice of data representation isn't as straightforward as Marr's example because neural networks learn to represent the data conditioned on high-level objectives obviating any fine-grained control over the representations. The design task has shifted from hand-crafting features to designing architectures and

training objectives that impose task-appropriate inductive biases. For instance, translating an image by a few pixels, which causes little to no change in the relevant constructs to be interpreted from the image, corresponds to a huge change in the pixel space. Therefore, an invariance to geometric transformations is a very useful inductive bias, and architectural choices like convolutional layers early in the network allow for the convenient imposition of such biases. The training objective is another crucial ingredient in the recipe as neural networks are infamous for learning only what they are incentivized to learn to the point of exploiting artifacts in the dataset and taking shortcuts to achieve the objectives [26]. There are inherent biases associated with particular objectives. For instance, minimizing cross entropy, a common objective for supervised classification problems, encourages the network to stop learning once simple predictors have been found [26]. Hence, the architecture and the training objective act as control knobs that allow us to coerce the representations in a direction that disentangles and makes explicit all the task-relevant information from the data.

2.2 Self-supervised learning

While most of the success of deep learning has been predicated on using neural networks as function approximators trained on expertly curated inputs and outputs i.e. supervised learning, the approach has some major limitations; manual annotation of data is time-consuming, expensive, prone to human biases, and subject to diminishing returns due to roughly logarithmic increases in performance upon the addition of data [27]. In recent years, self-supervised learning has emerged as a way to alleviate said limitations. Self-supervision eliminates the need for manual data annotation by generating labels algorithmically from the data itself, often leveraging the underlying structure in the data. *Self-supervision* refers to learning tasks that ask a model to predict one part of the input data—or a label programmatically derivable thereof—given another part of the input [28]. For instance, in natural language processing, parts of sentences are hidden and a model is tasked with predicting the hidden words with the rest of the sentence as the input. Similarly, predicting hidden patches in images has been used as a pretext task in computer vision.

While self-supervised learning doesn't require manual data annotation, a supervisory signal is still necessary. **Pretext tasks** are used as a way to generate the supervisory signal from the data. There are four main classes of pretext tasks - masked prediction, transformation prediction, instance discrimination, and clustering. In masked prediction, parts of the input are hidden or masked, and the training objective is to predict the masked input using the rest. This is the main workhorse of natural language processing being used exclusively or in part by most state-of-the-art methods [29-31]. And while it has shown some promise in computer vision as well [32], the continuous and high-dimensional state space of visual inputs, as opposed to a discrete and well-defined vocabulary in a language, poses a significant challenge. Transformation prediction has been proposed as a task in computer vision where given an image and a transformed version of it, the model is tasked with predicting what transformation has been applied and to what extent. For instance, given an image and a rotated version of it, the model would have to predict the angle of rotation [33].

The instance discrimination-based methods treat each instance in the dataset as its own class, and the training task is to discriminate between the instances. While attempts have been made to use the naive approach of treating this as a multiclass classification problem [34], a bulk of the success has come from using **contrastive learning**. In contrastive learning, the training task is to predict whether a pair of inputs belong to the same class instead of predicting the class itself. This makes it a binary classification problem making it much more tractable. The pair of inputs could be the original and augmented versions of an image, an image, and audio from the same video, an image, and a caption used alongside on a social media website, etc. In order to solve this problem, the representations of corresponding inputs have to be the same or be as close as possible. Optimizing for this objective leads to similar inputs being pushed close in the embedding space while non-similar inputs are driven further apart. It should be noted that the notion of similarity here is defined by the pretext task and the choice of data augmentations. Clustering-based methods are much more robust to the choice of augmentations and focus on dividing the training data into a number of groups with high-intragroup and low-intergroup similarity.

2.3 Sound source localization algorithms

This section compares state-of-the-art algorithms for visual sound source localization based on their training objective, localization method, architecture, and the utilization of temporal information.

2.3.1 Audiovisual correspondence

The natural correspondence between audio and images in videos has been leveraged as a supervision signal for self-supervised learning. Arandjelovic et.al. [9] introduced the audiovisual correspondence (AVC) task as a method for training audio-visual representations. It is framed as a binary classification task with the objective of predicting if an image-audio pair corresponds i.e. they both come from the same video. Separate vision and audio encoders are trained to solve the AVC task and the authors demonstrate competitive performance of their learned embeddings on a variety of audio and vision tasks (Fig.1 a). In a subsequent publication [10], the authors expand on the AVC task tailoring it for cross-modal retrieval (Fig.1. b), and in the process, making it effective for sound source localization as well. To retrieve audio that corresponds to a query video or vice versa, the audio and visual embeddings are made to have the same dimensions allowing the use of common distance metrics. The correspondence prediction is then made based on the Euclidean distance between the two embeddings essentially forcing positive pairs to be closer and the negative pairs farther apart which bears resemblance to contrastive learning. For localizing sound sources (Fig.1. c), the distance between audio and image embeddings is not calculated globally but is done in patches with cosine similarity instead of euclidean distance as the distance metric. The question in cross-modal retrieval is whether any region in the image highly corresponds to the audio. In sound source localization, however, the problem is reframed to also find out which regions correspond to the audio. The correspondence, in this case, is measured using the cosine distance, which is essentially calculating the correlation between the audio and the local patch-wise visual embeddings. This is based on the assumption that if the embeddings are properly trained, the regions of the image containing the sound source will be correlated with the audio. The correlation between

the audio and video of a car, for instance, will have high values for regions containing cars and much lower values for the regions in the background.

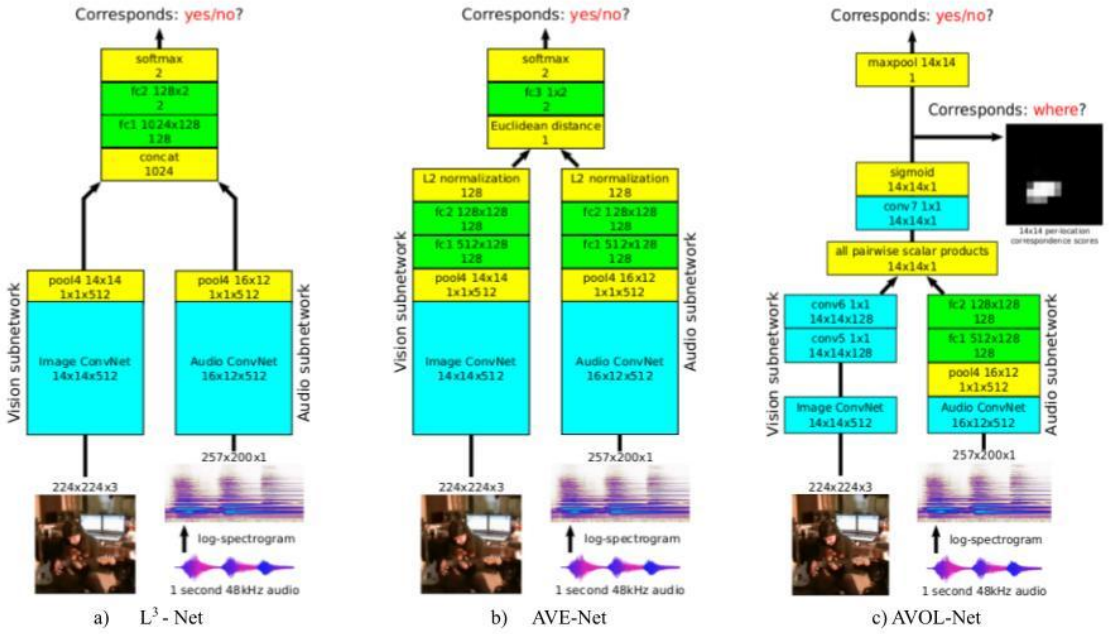

Figure 1. L[3]-Net [9] and the modifications made to adapt it for cross-modal retrieval and sound source localization [10]

Senocak et.al. [11] also use audio-visual correspondence with contrastive learning to train audio and vision embeddings. The approach is very similar to [10] but their primary objective is localizing sound sources instead of cross-modal retrieval. They use AVC in conjunction with a supervised loss that comes from labeled localization data making their method semi-supervised. Takashi et.al. [12] also employ AVC and contrastive learning as the training objective but they decompose the problem of localizing sound sources into two steps; first, they generate candidates for potential sound sources using the image only and then use the audio to filter the generated candidates. The candidate generation is done by using the activations of a VGG network pre-trained on ImageNet. The pre-trained embeddings impose an object-ness prior of sorts since ImageNet is an object-centric dataset and it is reasonable to assume that regions of images containing objects will have higher activations in embeddings trained

on ImageNet. The filtering is based on the similarity between audio and vision embeddings much like the other aforementioned methods. They also try to decouple the contribution from the two modalities - audio and vision - and find out that in many cases, especially with single large objects in the image, sound source localization shows very little improvement upon the inclusion of audio; hence the title of the paper "*Do we need sound for sound source localization?*". This reflects not on the problem of sound source localization in general but on the dataset used in the study. If a large proportion of the data contains single objects, then just identifying the object suffices to localize the sounding object despite the lack of explicit sound source localization in any real sense. This serves as a bit of a cautionary tale on how not acknowledging the biases in your data can engender inferences that don't generalize beyond the confines of the dataset.

Chen et.al. [13] propose an enhancement over the framework of using AVC with contrastive learning where they automatically mine for hard-negative examples for training. They divide the image into three regions - foreground i.e. the object that is making the sound, the background, and a small region of uncertainty between them. The foreground coupled with the audio from the same video is considered the positive pair and the background region with the audio from the same video is considered the hard-negative pair. However, as their method is self-supervised, they don't have these regions annotated, to begin with. In the absence of ground truth annotations, the model predicts masks by thresholding the cosine similarity between the audio and vision embeddings, and the mask is then used to generate the positive and negative examples for jointly training the audio and vision encoders. The training is done in a standard contrastive learning fashion - the embeddings for the positive pairs are pushed closer together while the ones for negative pairs are driven farther apart.

While [13] inherits the idea of positive and negative pairs for contrastive learning from its predecessors, albeit proposing a significant extension, Song et.al. [35] do away with negatives altogether and use a Siamese framework [36] to train audio-visual embeddings for sound source localization. They calculate an audio-visual embedding (f_{av}) for an audio-image pair by fusing the audio and visual features. They use the patch-wise cosine similarity to fuse the audio and visual features. f_{av} is obtained by

calculating a weighted average over all patches of the visual features using the similarity map as the weights. The similarity map hence acts as an attention mechanism to weigh the original visual features. The training is done in a typical Siamese fashion where the distance between the f_{av} for an image and an augmented version of it with the same audio is minimized. Additionally, they propose a predictive coding module, which is implemented using a recurrent neural network, to align the audio and vision embeddings by predicting one using the other.

Building on the idea of inferring objects from images as a precursor step to sound source localization as in [12], Mo et.al. [14] propose biasing the localization towards objects present in the image by introducing an object encoding. They use the activations of a resnet-18 network pretrained on Imagenet[37] as the object encoding with the assumption that pretraining on an object-centric dataset would result in high activation in regions of the image containing objects. An intermediate localization map is obtained by calculating the region-wise cosine similarity between the audio and image embeddings as in [13]. The final localization map is a linear combination of the object encoding and the cosine similarity map biasing the localization towards objects. Another novelty of this method is the training objective; it uses AVC with multiple-instance contrastive learning where the audio embeddings have to be similar to at least one region in the image. In contrast to [13] where the loss function is calculated using the average similarity over a region, [14] uses the maximum of similarity values across the image. The authors compare the performance to the best performing methods [10, 13] on VGG-SS and Flickr-Soundnet and demonstrate that their method significantly outperforms the competition making it the current state-of-the-art.

2.3.2. Beyond audiovisual correspondence

While the correspondence between audio and images has proven to be very effective for learning representations, it is by no means the only possible source of supervision. Classification has been used as a supervision signal in conjunction with AVC by Qian et.al. [38] to perform sound source localization for the case of multiple sources. They separate a complex audio-visual scene into several simple scenes with multi-class classification and then use class activation mapping (CAM) to disentangle the different sound sources and predict localization maps for each. Senocak et.al. [39] eliminate AVC

altogether and perform sound source localization using only classification. They propose a network with separate audio and vision encoders followed by 3 classification heads - audio, visual, and audiovisual. The audiovisual features are obtained by simply concatenating audio and visual features and the training minimizes the cumulative classification loss across all modalities. Following the training, sound localization is performed using the cosine similarity between the audio and visual features. One key difference between [39] and the other methods discussed so far is the inclusion of temporal information from the video achieved by using 3D convolutions in the early layers of the vision encoder resulting in spatio-temporal visual embeddings.

The temporal context is more explicitly utilized by incorporating it in the training objective by extending AVC to audiovisual synchrony [40, 41]. Korbar et.al. [40] show that the synchrony between audio and images in videos can be used to train representations. They define the task under a contrastive learning framework where positive examples are synchronized audio-video pairs whereas N-shifted versions, i.e. audio-video pairs where either has been shifted by a small time duration, are considered negative examples. This is a stricter condition than AVC and the model is forced to learn the temporal correlation between audio and video in addition to the semantic information. Afouras et.al. [41] use this task for sound source localization where they transform videos into a set of discrete audio-visual objects. They also use 3D convolutions in the vision encoder to imbue the visual embeddings with temporal information. To that end, they also aggregate cosine similarity-based attention over time using optical flow followed by peak finding and non-max suppression (NMS) to extract trajectories for each audiovisual object. They demonstrate competitive performance on sound source localization with multiple sources i.e. speaker detection and tracking in a talking heads scenario with multiple speakers. This is a significant enhancement over AVC-based methods as it wouldn’t be possible to distinguish between speakers just by using similarity between semantic audio and vision embeddings as the metric for localization.

Another class of models that are quite distinct from the ones discussed so far use a teacher-student architecture where knowledge from vision or other modalities (teachers) is distilled into an audio network (student) [42-44]. There are several key differences;

the localization is performed using only the audio, the audio used is stereo or multi-channel whereas all the aforementioned methods work with mono audio, and there is knowledge distillation into the audio network from other modalities. Chuang et.al [42] use a pretrained YOLOv2 object detection model as the teacher network and distill its knowledge into an audio network that takes stereo audio as input and regresses bounding boxes for the sound source as an output. After training, the model is able to detect moving vehicles using only the audio. However, it is not clear how the model gets any information about the vertical coordinates of the vehicles from stereo audio and it is possible that it is overfitting with biased priors on the average size and vertical position of vehicles in the dataset. Valverde et.al. [43] also perform vehicle detection using teacher networks trained on vision, depth, and thermal data and distilling them into an audio network that takes 8-channel audio as input. Zürn et.al. [44] choose moving vehicles as their objects of interest as well but they train an audio-visual teacher using AVC and contrastive learning and subsequently distill it into an audio-only network. They perform experiments with multichannel audio with 2, 4, and 6 channels and notice very small differences in performance upon increasing the number of audio channels with no clear trend as to how changing the number of channels affects localization.

2.3.3 Limitations

The state-of-the-art methods and the datasets used to develop them have the following limitations. Most methods do not use any temporal context and the audio-visual embeddings are purely semantic. Moreover, as shown in [15], patch-wise cosine similarity as a localization method results in large and diffused sound source estimations. Such localization works well for commonly used benchmark datasets because of the prevalent bias of having single large objects in the middle of the image. However, this method falls short for urban scenes where it is common to have many sound sources spread across the image. This thesis attempts to address these limitations by introducing motion information into the state-of-the-art localization method proposed in [15] that does away with cosine similarity and uses explainability techniques for localizing sounding objects.

Chapter 3

Methodology

3.1. Urbansas dataset

Urban Sound & Sight [16] (Urbansas) is an audio-visual dataset developed for studying the detection and localization of sounding vehicles in the wild. The dataset consists of labeled and unlabeled videos of urban traffic with stereo audio. The videos are sourced from two publicly available datasets - TAU Urban Audio-Visual Scenes 2021 Development dataset [45] and the Montevideo Audio-Visual Dataset (MAVD) [46]. TAU is a general-purpose audio-visual dataset and only the subset containing traffic videos has been included in Urbansas. MAVD is an audio-visual traffic dataset annotated with vehicle sounds and is intended for sound event detection. The two sources add up to a total of 15 hours of video out of which 3 hours have been manually annotated, with both audio and video annotations, for sound event detection and source localization.

3.2. Evaluation metrics

The evaluation of visual sound source localization methods is done using some variant of the Jaccard index i.e. intersection over union (IoU). IoU has been adapted from computer vision where it is the *de facto* metric to evaluate object detection, segmentation, and tracking models [47]. Given a ground truth (A) and a prediction (B), IoU is defined as follows.

$$IoU = \frac{A \cap B}{A \cup B}$$

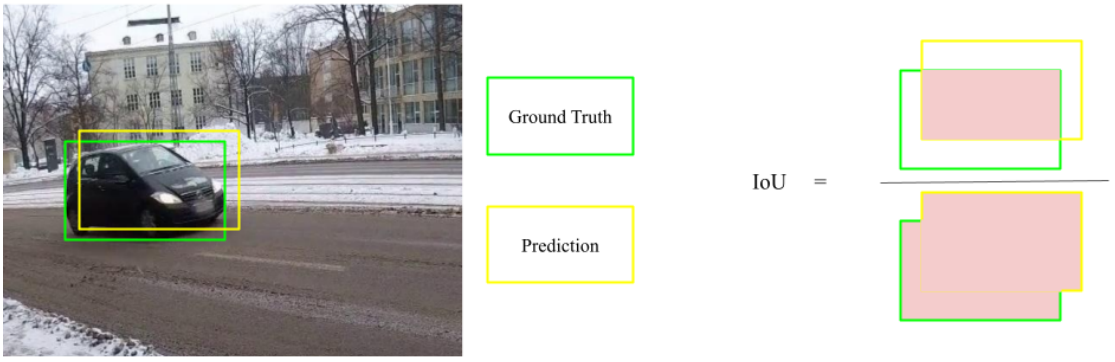

Fig 2. IoU is defined as the area of intersection over the area of union between the ground truth and the prediction made by a model

IoU has certain desirable properties like invariance to scale, computational efficiency, and adaptability to arbitrarily shaped annotations and predictions; however, it comes with its own set of limitations. One major drawback of IoU is that once there is no overlap between the ground truth and the prediction, it cannot distinguish between different predictions. While fig 3.b is clearly a much worse prediction than fig.3.c, the IoU for both cases is 0. Moreover, if IoU is only used only as an evaluation metric and not as a loss for training, there can be a disassociation between training and evaluation objectives. For instance, if a model is trained to predict precise segmentation maps while it is evaluated by calculating IoU with bounding boxes, there will be a significant deflation in performance and the metric would become subject to irrelevant features of the data like the orientation of objects.

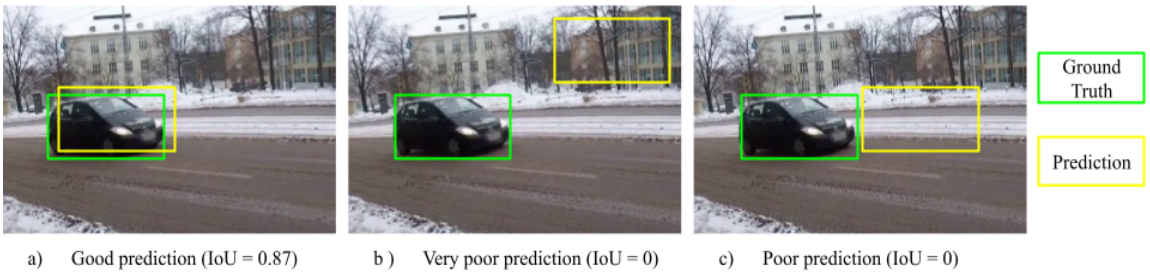

Fig 3. Examples of IoU demonstrating that it can not distinguish between predictions that have no overlap with the ground truth

3.3. Baselines

3.3.1 Vision-only baseline

We have implemented a vision-only baseline using a pre-trained YOLOR object detection model [48]. The yolor_p6 model weights provided in the official repository[1] were used for inference, and the results were filtered to four vehicle classes - car, motorcycle, bus, and truck. Further motion-based filtering was applied where motion is used as a heuristic for sound with the assumption that moving objects are also sounding objects. For each pair of consecutive frames (*f* and *f +1*), if a bounding box in *f* has an IoU greater than 0.95 with one in *f+1*, both the bounding boxes are discarded. This ensures that stationary objects are filtered out in the final predictions. The remaining predictions are evaluated against the ground-truth labels in urbansas. This method is supposed to only serve as a baseline since audio information is not considered in localizing sounding objects. It serves to demonstrate the correspondence, or a lack thereof, between moving and sounding objects.

3.3.2 RC-Grad

RC-Grad [15], the current state of the art on visual sound source localization for Urbansas, has been used as an audio-visual baseline. We have replicated the results of [15] using the pretrained models provided by the authors in the official repository[2]. As is standard in the literature, the model uses separate audio and vision encoders optimized with audio-visual correspondence as the training objective. RC-Grad uses resnet-18 as the audio as well as the vision encoder. The vision encoder is pretrained on Imagenet while the audio encoder is randomly initialized. The model is then trained with a contrastive loss on VGG-Sound [49].

Grad-CAM [50] has been used as the method for localization. Grad-CAM is an explainability technique that uses the gradients of a target concept (say ‘dog’ in a classification network) flowing into the final convolutional layer to produce a localization map highlighting the important regions in the image for predicting the concept. For sound source localization, instead of backpropagating the classification

[1] YOLOR [48] - https://github.com/WongKinYiu/yolor
[2] RC-Grad [15] - https://github.com/rrrajjjj/rethinking-visual-sound-localization

output, the audio embedding itself is back-propagated through the vision encoder. For evaluation, 1 second of audio and the image for the frame corresponding to the middle of the audio segment is used as input to the model.

3.4. Incorporating temporal context

3.4.1 Optical flow

Optical flow is the pattern of the apparent motion of objects in a visual scene caused by the motion of an object or camera or both [51]. The origins of the idea can be traced back to animal psychophysics [52] which inspired the mathematical formalism in Horn's seminal paper [53] and his subsequent book *Robot Vision* [54]. In computer vision, optical flow is a vector field that, given two consecutive frames of a video, tells how much each pixel has moved and in what direction (Fig 4). In the 4 decades since [53] a plethora of methods, both traditional [53-58] and data-driven [59-62], have been proposed to estimate optical flow. Most traditional methods exploit assumptions like the constancy of intensity of each pixel, homogeneous illumination, similar flow values for neighboring pixels, etc. The more modern data-driven approaches are based on neural networks and learn to estimate optical flow with large amounts of ground-truth training data.

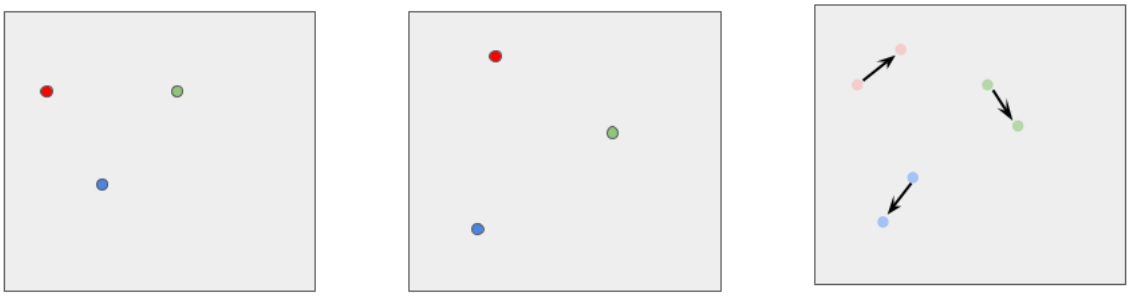

Fig 4. *Optical flow encodes the movement of pixels across images*

In this work, the OpenCV implementation of the Gunnar Farneback algorithm [58] is used to estimate dense optical flow. Videos are sampled at 8 frames per second and are

converted to grayscale before calculating the optical flow. The flow vectors are then converted from cartesian to polar coordinates and only the magnitude is used in subsequent analysis. Out of the 3 available frame-rates in the dataset, 8 fps was selected based on empirical evaluation on a group of videos; an example can be seen in Fig. 5. At 2 frames per second, there is a significant change between the subsequent frames and optical flow operates on the assumption of small changes between images. It can be seen that the flow estimation is smeared out and does not have an adequate spatial resolution in terms of localizing the moving vehicle. At 24 frames per second, the change between the two frames is so small that the optical flow values of the vehicle are comparable to the background and hence it is not very helpful in gleaning the motion of interest from the video. 8 frames per second works well at precisely locating the vehicle while also ensuring that the flow values for the vehicle are significantly higher than that of the background making it the most suitable sampling rate for this dataset.

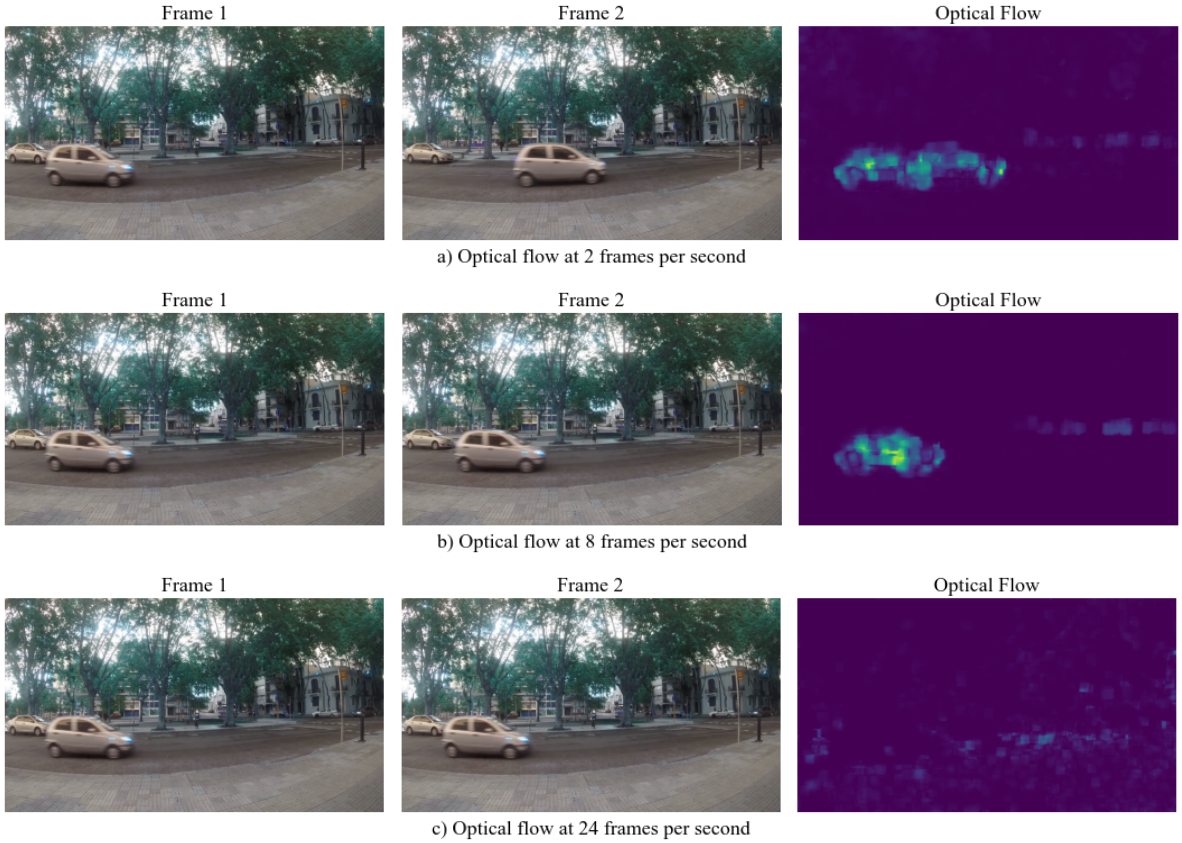

Fig 5. *Optical flow calculated on a video at different frame rates. a) 2 fps b) 8 fps c) 24 fps*

3.4.2 Justifying the use of optical flow

Optical flow has been applications in a wide range of computer vision tasks; it has been used, either as a primary feature or in conjunction with other visual features, in visual surveillance [63], object segmentation and tracking [64], action recognition [65, 66], pose estimation image superresolution [67], and deepfake detection [68]. It has also been used for sound source localization in [41] where image-wise localization maps are aggregated over time using optical flow to get coherent object trajectories. The overarching rationale behind the above-mentioned methods is that information about the motion of objects is crucial to visual perception and can be leveraged to great effect. This is especially true if moving objects are of interest. The authors of [69] go as far as using only[3] optical flow as a feature for camouflaged object detection and achieving state-of-the-art performance despite not utilizing the rich RGB images as inputs which is conventional in the deep-learning era.

In the context of Urbansas, where the goal is to localize the sounding objects in a video, motion is a very potent indicator of sound. In the urban traffic setting, the sounding object is in most cases also a moving object. Moreover, looking at single images devoid of motion information, it can often be impossible to detect the sounding objects. For instance, parked cars are indistinguishable from moving cars given just an image. In Fig.6, it can be seen that RC-Grad mistakes the large parked car as the primary sound source. However, simply thresholding the optical flow generates a prediction that is very close to the ground truth.

Fig 6. Comparison of optical flow and RC-Grad at localizing the sounding vehicle

[3] They also use the difference between the frames after correcting for camera movements to recover sharp object boundaries. The use of "only" is supposed to imply the authors discarding the original RGB images as inputs to their deep neural networks where end-to-end learning has become the norm.

While *Fig.6* clearly demonstrates an example where optical flow helps overcome a big limitation of state-of-the-art sound source localization models, inferences made on cherry-picked examples have to be taken with a grain of salt. Generalisability cannot be assumed given the heterogeneity and complexity of the dataset. Therefore, to complement the above justification for using optical flow, a statistical comparison between optical flow and ground truth annotations has been done. *Fig 7. a* shows the distribution of ground truth annotations i.e. bounding boxes and *Fig 7. b* shows the aggregate of optical flow both calculated over the entire dataset divided by cities. An aggregate over the entire dataset has also been reported.

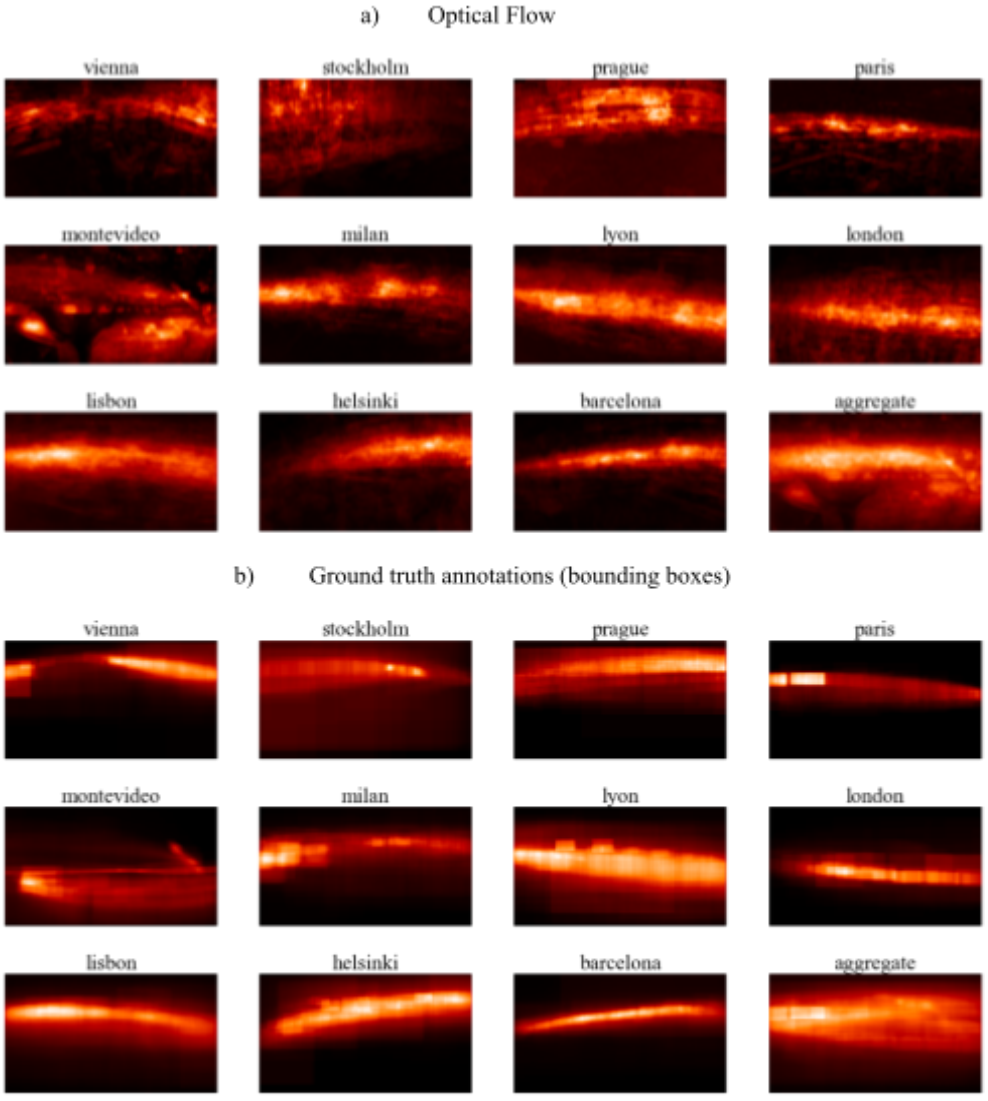

Figure 7. Density distribution of optical flow and bounding boxes aggregated over cities

It can be seen that the distribution of optical flow very closely resembles that of the bounding boxes corresponding to sounding objects. While this doesn't directly imply the coincidence of moving and sounding objects, the almost identical spatial distributions suggest a meaningful relationship between the two. The noticeable differences in Stockholm and Montevideo however need to be explained. In the case of Montevideo, most videos have large trees occluding the view of the street and the movement of leaves, which happen to be closer to the camera than the vehicles, is significant enough to be picked up by optical flow (Fig 8. a). The disparity in the case of Stockholm can be attributed, at least in part, to a non-stationary camera (Fig 8. b). One of the assumptions in using optical flow to detect moving objects is that the camera is stationary so that only moving objects in the scene contribute to the optical flow. When there is global movement due to a shaky camera, all points in the image have high values of optical flow. Also, the optical flow for objects in the vicinity of the camera is much higher than it is for objects that are far away due to the parallax effect. Moreover, there is a higher number of pedestrians in the videos from Stockholm. Since the movement of pedestrians is also picked up by optical flow but they are not labeled as sounding objects in the dataset, they also contribute to the disparity between moving and sounding objects.

a) Montevideo - high values of optical flow despite the absence of moving vehicles due to the movement of leaves

b) Stockholm - high values of optical flow due to camera movement. The closer the object, the higher the optical flow.

Figure 8. Examples demonstrating the limitations of optical flow in the context of Urbansas

Despite the aforementioned caveats, the clip-level example from Fig.6 along with the more statistical dataset-level comparison between moving and sounding objects justifies the use of optical flow as a useful feature for visual sound source localization at least under the confines of Urbansas.

3.4.3 Optical Flow as a heuristic

One significant limitation of RC-Grad, and most other state-of-the-art methods, is parked vehicles that don't make any sound. Since the representations are purely semantic and there is no temporal context, the model cannot distinguish between stationary and moving vehicles. As a result, parked vehicles often end up having high activations as false positives diminishing the performance (Fig.9). Optical Flow, on the other hand, only has motion information. Anything that moves, be it vehicles, people, tree leaves, etc., have high values of activation. Hence, optical flow and RC-Grad have complementary strengths that can be leveraged by taking an intersection of objects that have high activations for both (Fig.9). Here, that has been implemented by a simple element-wise multiplication of the RC-Grad predictions for an image with the optical flow between the image and the next frame of the video. This suppresses objects that are either not moving or not vehicles leaving us with moving vehicles which is very much in the spirit of the vision baseline but has the added audio component.

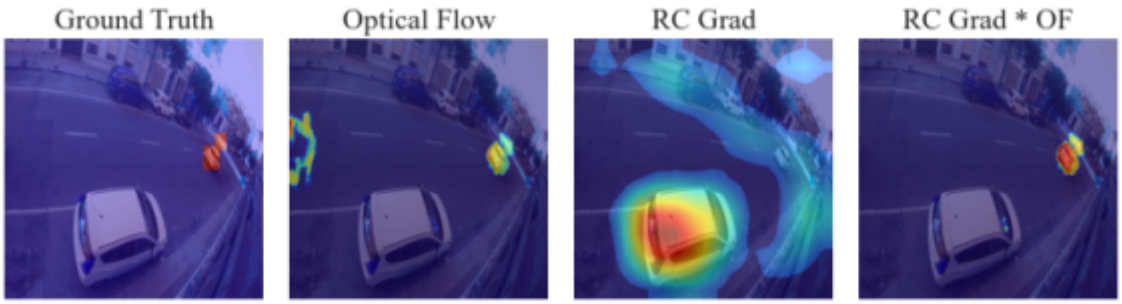

Figure 9. Optical flow as a heuristic to filter out stationary vehicles

3.4.4 Optical flow as an image channel

As effective as heuristics can prove to be, they are often rigid, brittle, and prone to a lack of generalizability. In an attempt to move away from the naive use of optical flow as a filter and towards using it to imbue the representations with temporality, we include

it as an image channel. Here, the model can, at least in principle, take the motion information into account while making predictions instead of motion being used as a filter post-hoc. The relationship between motion and sounds can hence be learned.

RC-Grad has been extended to take in 4 channels (RGB and optical flow) as the input to the image encoder (Fig.10.a). The model is initialized with the pre-trained version of RC-Grad. The weights of the optical flow channel have been initialized by averaging the weights of the RGB channels. The model is then trained on the unlabeled portion of Urbansas. The model is trained similarly to [15] where a video frame with its corresponding optical flow as an additional channel and a 5-second audio clip around the video frame has been used as inputs. The optical flow is calculated using the OpenCV implementation of the Gunnar Farneback algorithm.

3.4.5 Optical flow encoder

Optical flow and RGB images are distinct data modalities and in the above-mentioned method, the 4 channels of the image are pooled in very early layers of the network. This may result in shallow integration of motion information. Moreover, since the model was initialized with weights pretrained on audio and images, simply discarding the additional optical flow channel provides a trivial solution to quickly minimize the loss which, while making the experiment redundant, is a very real possibility. As a method to overcome said issues, a separate flow encoder with the same Resnet-18 architecture as the image and the audio encoders has been added to RC-Grad. The weights are initialized as the average of RGB channels of the vision encoder. The training loss has been modified to be the sum of all pairwise losses (audio-image, image-flow, and audio-flow). The localization is then done by backpropagating the audio embeddings through the image as well as the flow encoder to generate two localization maps. These maps are then multiplied element-wise to give the final localization map.

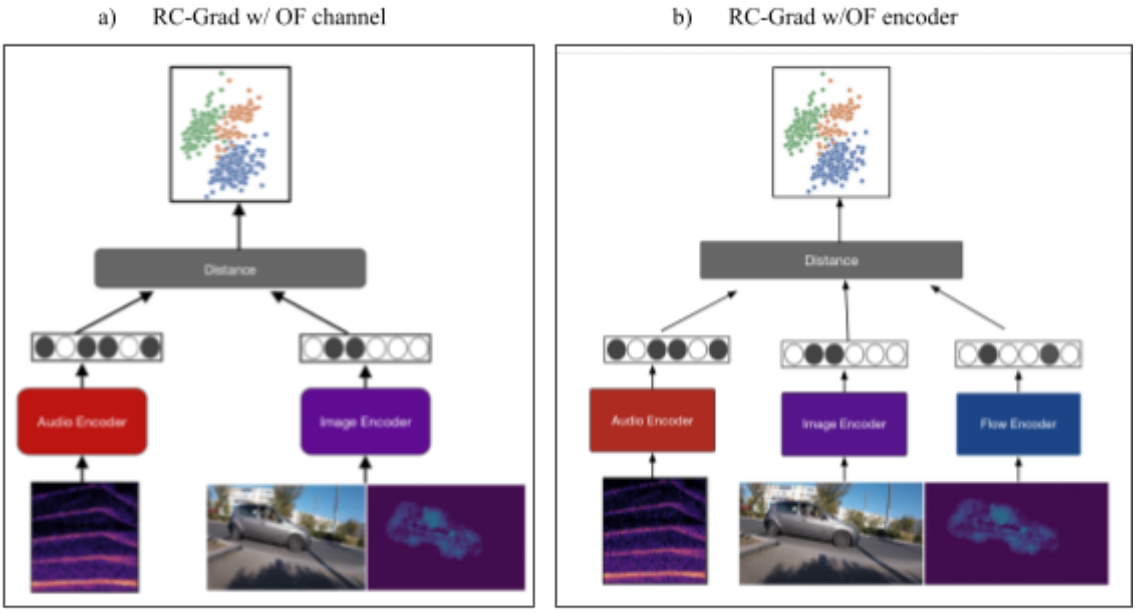

Fig 10. Architectures of extensions to RC-Grad to include optical flow as an input

Chapter 4

Results

The vision baseline significantly outperforms RC-Grad (Table.1) which can be attributed to several factors. The object detection model has been trained to predict precise bounding boxes around objects on the MS-COCO dataset [70], which is a large-scale dataset with just under a million annotated objects, nearly 10% of which correspond to vehicles. Further, stationary vehicles are eliminated with the motion-based filtering helping overcome a major limitation of RC-Grad. Moreover, there is a congruence between the predictions of the vision baseline and the ground truth annotations i.e. both are binary bounding boxes. RC-Grad predicts localization maps that have a continuous range of values where the higher the values, the more the region corresponds to the sound source. These maps are thresholded to generate arbitrarily shaped binary masks that are then compared against the ground truth annotations for evaluation. This mismatch most likely deflates the evaluation metric resulting in overall lower performance for RC-Grad.

Table 1. Performance on Urbansas

Model	**IoU**	**AUC**
Vision-only Baseline	0.32	0.21
RCGrad	0.16	0.13
Optical Flow only	0.33	0.23
RCGrad * Flow	**0.50**	**0.30**
RCGrad w/extended vision encoder	0.26	0.18
RCGrad w/Flow Encoder	0.37	0.23

4.1. Optical Flow as a heuristic

Optical flow significantly improves performance over vanilla RC-Grad. As can be seen in Fig.11, it helps overcome a major limitation of RC-grad and other state-of-the-art methods - parked vehicles. In all three cases in Fig.11, the predictions of RC-Grad center on parked cars. These vehicles are not generating any sounds and hence are false positives. However, since they are stationary, they are not picked up by optical flow, and by multiplying the predictions with optical flow, these vehicles are filtered out.

To get a more well-rounded picture of the performance, the distribution of image-wise IoU values has been plotted for RC-Grad, optical flow, and the product of both across different lighting conditions and data sources (Fig.14). It can be observed that multiplying the predictions of RC-grad with optical flow shows improvement over both across all conditions as the distribution shifts towards the positive end. Fig.13 shows how multiplying the two improves or degrades performance at the level of individual images. In the left figure, it can be seen that most points lie above the $x=y$ line, which means that there is an improvement in most cases. In the right figure, there is a similar trend, however not as pronounced, indicating that RC-Grad contributes positively to optical flow.

However, there are points below the line as well (Fig.13) suggesting that optical flow is not a silver bullet and has its drawbacks. There are cases where multiplying the predictions with optical flow degrades performance. In Fig.12 a), the vehicle stops at a signal while keeping the engine on and hence is generating sound. But since it is stationary, it is filtered out resulting in a false negative. In Fig.12 b) and c), RC-Grad is able to localize the sound source precisely. Optical flow however does not perform well in these cases due to a shaky camera and when paired with the RC-Grad predictions, degrades the overall prediction quality.

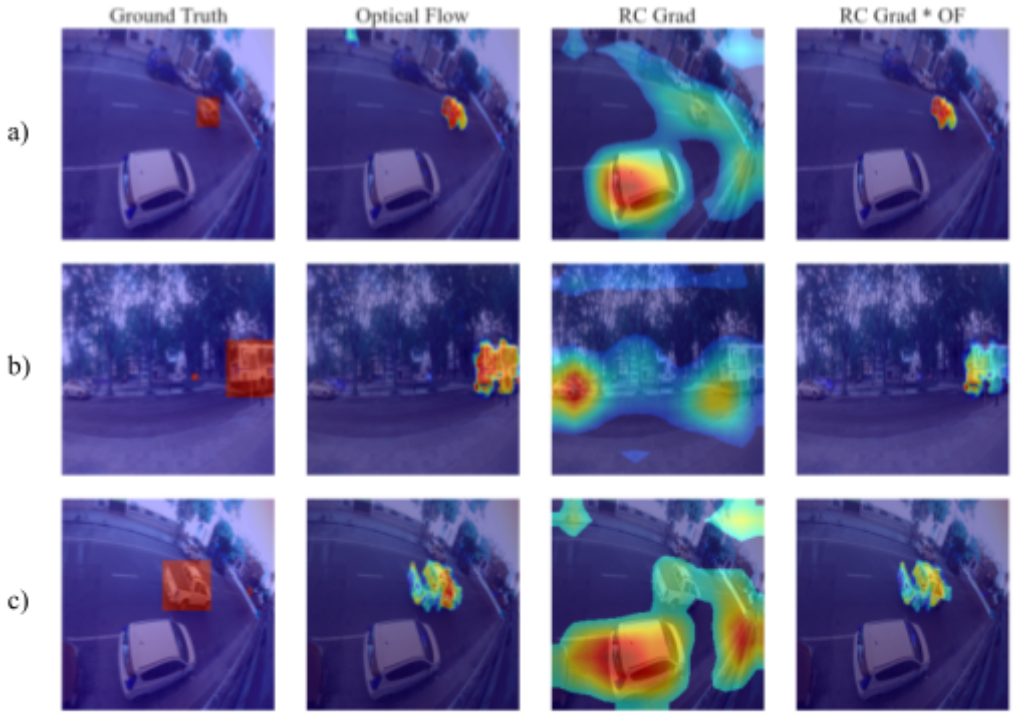

Fig 11. Examples where optical flow improves performance over RC-Grad

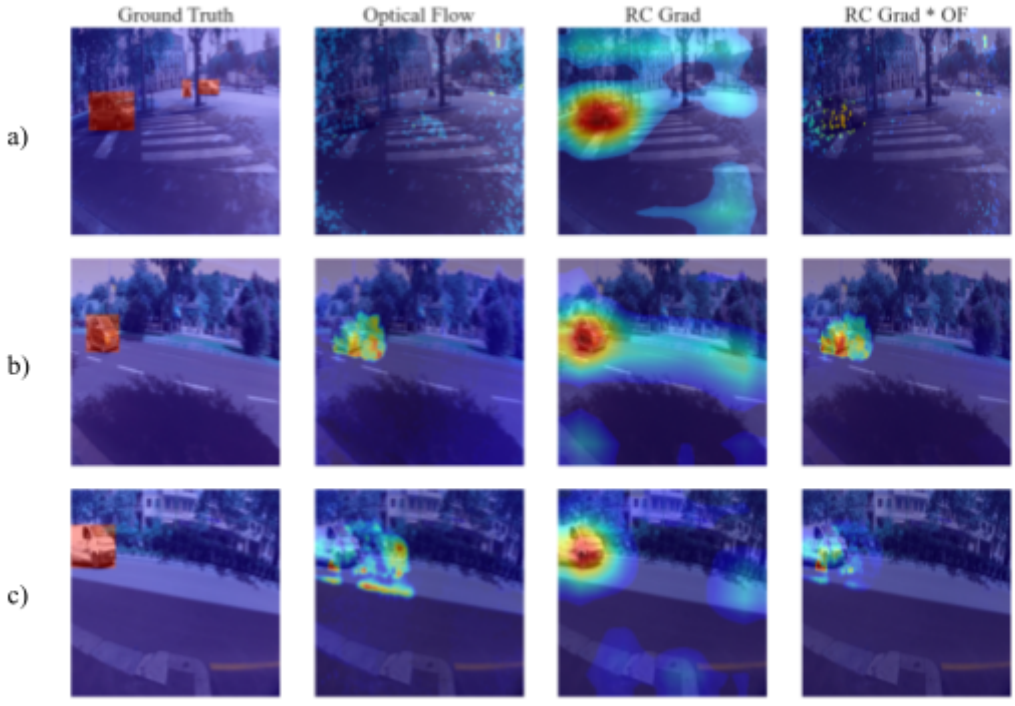

Fig 12. Examples where optical flow degrades performance over RC-Grad

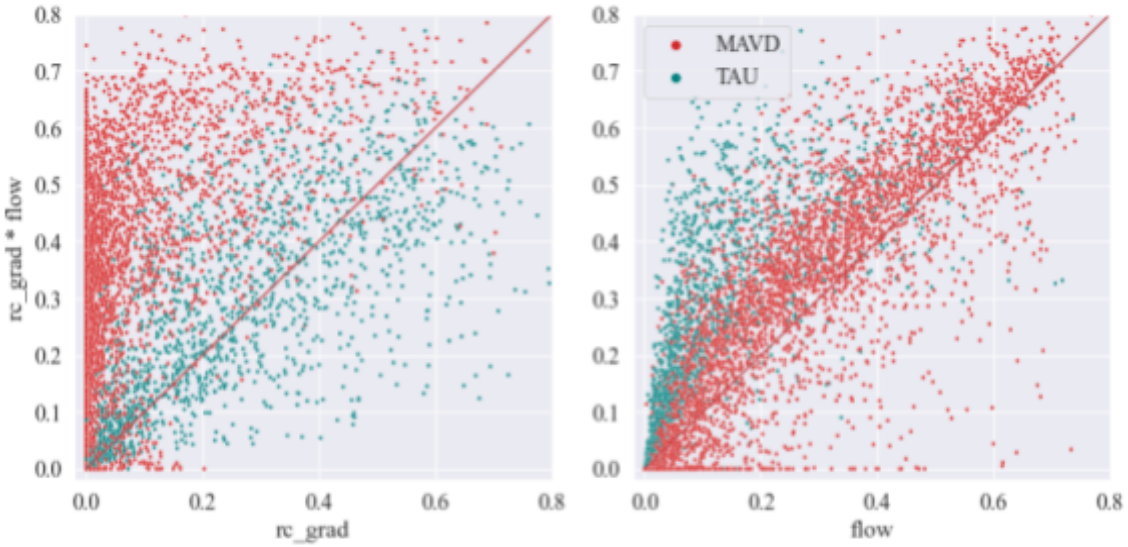

Figure 13. Image-wise effects of multiplying optical flow with RC-Grad predictions

The possible degradation of performance upon the introduction of optical flow motivates an analysis of factors that may affect optical flow estimation. The following factors have been considered -

1. **Lighting -** Lighting conditions may affect the performance of not only optical flow but also RC-Grad. The effects of lighting have been assessed using the day-night annotations in Urbansas (Fig.14 a).

2. **Non-stationary camera** - An unstable camera breaks an underlying assumption that things that move generate sound. If there is global movement, all regions in an image have high values of optical flow and the quality of flow estimation for moving vehicles is also compromised. There is no straightforward way to break down the performance of the models subject to camera stability since such annotations are not available. However, an unstable camera has been observed exclusively in the subset of Urbansas that comes from TAU. Hence, the data source has been used as a proxy for assessing the impacts of a shaky camera (Fig.14 b).

3. **Speed -** The strong effects of the frame rate of a video on optical flow estimation (Fig.5) suggest that the rate of relative displacement of objects between frames influences optical flow. The speed of objects has been measured

by calculating the distance moved by the center of the bounding box between consecutive frames per unit time and has been reported in pixels per second.

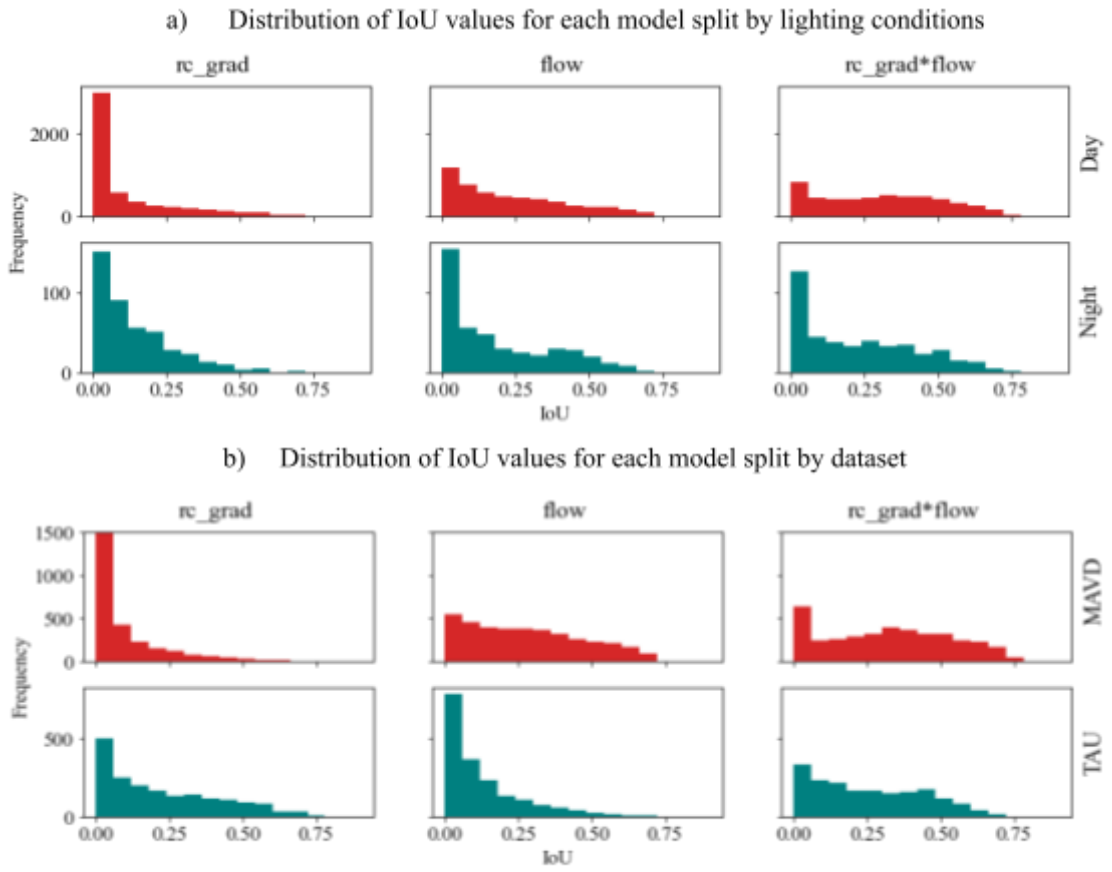

Figure 14. A breakdown of performance by lighting conditions and data source.

Poor lighting conditions have detrimental effects on the performance of optical flow as can be seen in Fig.14.a. There is a noticeable shift towards 0 in the distribution of IoU values for videos taken at night for both the models that rely on optical flow. RC-Grad on the other hand performs marginally better at night.

The data source, which is serving as a proxy for camera stability, also impacts performance. As we move from MAVD to TAU, the performance of optical flow suffers a significant drop (Fig.14.b center). The IoU values are concentrated around 0 for TAU but much more evenly distributed for MAVD. Also, Fig.13 shows that most cases where multiplying RC-Grad predictions with optical flow makes them worse, i.e. points that lie below the $x=y$ line, come from the TAU dataset. Manual inspection of these videos

shows a non-stationary camera in most cases. Out of the 15 such videos, only one shows a drop in performance due to the vehicle stopping at a signal.

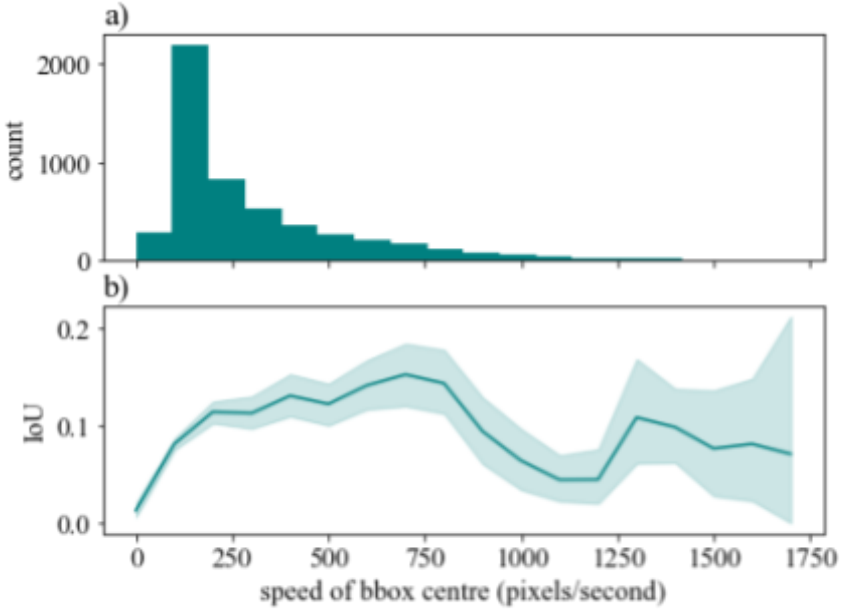

Figure 15. The effect of speed of objects on optical flow. a) the distribution of speeds across objects in the dataset. b) IoU vs speed of bounding box

The speed of the bounding boxes only minimally affects performance in the intermediate range (Fig.15). The IoU drops to 0 for very low speeds as is expected since stationary objects do not contribute to optical flow but looking at the distribution of the speeds reveals that most sounding vehicles are in the intermediate range with a strong bias towards the lower end.

4.2. Learning with optical low

Optical flow has been incorporated into the model in two ways - as an image channel and with a separate optical flow encoder. While both methods substantially outperform vanilla RC-Grad (*Table.1*) at an aggregate level, visualizing the predictions on selected test cases shows the results to be more of a mixed bag. Fig.16.a shows a marked improvement over RC-Grad as both the models learn to ignore the stationary car and infer the moving vehicle to be the sound source.

However, the model with optical flow as a channel inherits not only the strengths but also the weaknesses of optical flow. Fig.16.b and c have vehicles parked at a signal. The model predicts a diffused mask missing all the vehicles suggesting that predictions primarily rely on optical flow and the model goes awry in the absence of strong optical flow. Fig.16.d has a parked car on the left and a moving one on the right with the parked car being a false positive. The false positive is to a somewhat lesser extent for the model with optical flow as an image channel where the parked car has a marginally lower activation while the moving car has a more precise mask around it. The over-reliance of this method on optical flow can be tackled by training the model on more data where optical flow is not indicative of the sounding object or is completely absent. Videos with many pedestrians, cars parked at signals, or without any vehicles at all would be some scenarios with which to supplement the training data.

RC-Grad with a flow encoder performs much better than the rest in cases where clear optical flow information is available. It is able to eliminate the parked vehicle in Fig.16.d that shows up as a false positive in the other methods. The localization maps are also less diffused and this stringency is likely to contribute to the increased IoU numbers due to a decrease in the overall area of union. By the same token, the size of the predicted masks may also, at least in part, explain why this method doesn't perform as well as the naive use of optical flow as a heuristic. Optical flow generates very precise masks around objects minimizing the area of union and hence increasing the IoU while this method still produces diffused localization maps.

All the methods struggle in cases where optical flow is not reliable (Fig.16.b and c).. This motivates further investigation into better and more robust ways of estimating optical flow since the quality of optical flow is crucial to the performance of these methods.

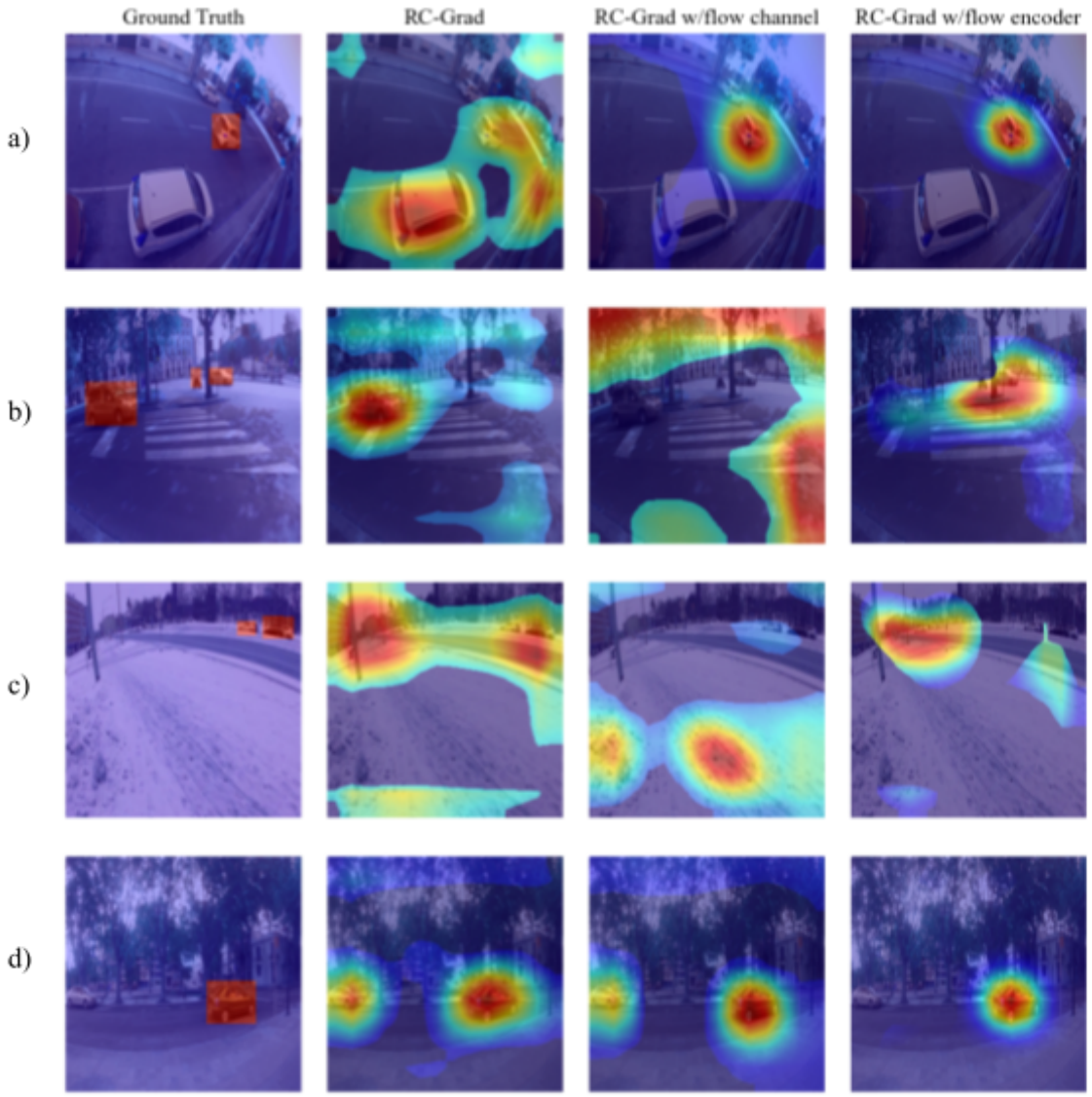

Figure 16. Predictions of RC-Grad and proposed extensions

Chapter 5

Discussion

This thesis tackles the problem of visual sound source localization in an urban setting with the main objective being the addition of temporal context to state-of-the-art methods. Urban traffic provides a more challenging scenario for localizing sounding objects than the commonly used benchmarks where there is a pervasive bias towards having large objects in the middle of the image [15]. The challenge is further amplified by the prevalence of multiple potential sound sources where it is not possible to infer which ones are active based on a single image. For instance, a moving car can not be distinguished from parked ones given a single frame motivating the use of context clues present in videos. Early studies in localizing sounding objects in videos relied primarily on the temporal correlations between video and audio [6] but recent works seem to completely disregard this source of information [10-14]. This is by design as addressed by the authors of *Objects that Sound* [10] where they suggest that models with access to temporal information can cheat by exploiting low-level information like motion. The goal of these methods is to learn semantic audio-visual representations and a model with access to temporal information has less incentive to do so.

The commonly used datasets have a large number of diverse classes and the models trained on these datasets essentially learn to predict whether a sound and an image, or a region therein, belong to the same class. Urban traffic data, however, is peculiar in two key ways - it has a very small number of unique classes but often many instances of each class in a single image. In such a case, semantic similarity devoid of temporal context does not get us very far. As mentioned earlier, parked vehicles pose a major limitation to this approach. However, the motion of objects between consecutive frames proves to be a strong indicator of sounding objects as we have demonstrated in our vision-only baseline. We have used optical flow to incorporate motion information into state-of-the-art sound source localization models and observed significant improvement in performance on Urbansas.

Optical flow is a remarkably effective feature for sound source localization within the context of Urbansas to an extent that simply thresholding optical flow to localize sounding objects outperforms the baselines (*Table.1*). In this work, we have explored three ways of using optical flow in conjunction with RC-Grad which is the state-of-the-art sound source localization model for Urbansas. The simplest method of multiplying the predictions of RC-Grad with optical flow and using the result as the localization map performs better than the more sophisticated approaches where optical flow is used as a feature and the relationship between motion and sound is learned. Moreover, this method allows convenient decoupling of the contributions of optical flow and RC-Grad enabling an analysis that reveals the strengths and limitations of optical flow based approaches for Urbansas.

In most cases, optical flow improves the localization of sound sources (*Fig.13*) by filtering out stationary vehicles and constraining the boundaries of the localization map to the moving object (*Fig.11*). However, there are limitations that persist across all the methods that will be discussed in the following section and avenues for future work will be suggested to overcome said limitations. It should be noted as an important caveat that in this study we only make inferences about the effectiveness of a method under the constraints of a specific dataset and we make no claims about the generalizability of said methods for the problem of visual sound source localization in general.

5.1 Limitations

5.1.1 Assumptions

The use of optical flow as a feature for localizing sound sources in videos is predicated on the following two assumptions -

1. Sounding objects are always moving objects
2. Optical flow faithfully represents the motion of objects of interest

The above two statements represent the ideal case and we must add further qualifications to adapt them to a real-world scenario. The first assumption can not be

true at all times, especially if we move outside of an urban traffic environment. A guitar amplifier on a stage for instance will not have any appreciable optical flow despite being a predominant sound source. Motion is not an unerring predictor of sound. However, we have observed that within Urbansas, sounding objects are very often moving objects. This is supported by our results (*Table.1*) where simply thresholding optical flow to localize sounding objects shows competitive performance. To restate the second assumption, we assume that optical flow perfectly captures all motion of interest i.e. all moving vehicles generate a strong optical flow signal. It is also to say that objects that are not of interest do not contribute significantly to optical flow. There is also an implicit assumption here that there is no global motion i.e. the camera is stationary and stable. When any of these assumptions are violated, the performance of our methods is severely impaired leading to the limitations discussed hereafter.

5.1.2. Limitations of the method

Vehicles parked at signals with their engines on violate the first assumption and, as demonstrated in *Fig.12.a.,* are a limitation of the method. Using short-term optical flow completely filters out such vehicles. Trees, pedestrians, and other moving objects are also exceptions to the assumption. Moving tree leaves can often have high optical flow as can be seen in *Fig.8.a.,* but they have no contribution to the sounds whatsoever. However, using optical flow along with RC-Grad is a simple fix to that issue as the predictions of RC-Grad generally have very low activations for trees.

5.1.2. Limitations of the dataset

The case with pedestrians is not as straightforward as it is for trees. They have characteristic sounds associated with them that are clearly audible, especially if they are close to the microphone. The models we use for sound source localization are class-agnostic and are trained in a self-supervised manner without any class labels. So RC-Grad learns to localize pedestrians as sound sources as we have observed in some cases (*Fig.17*). Pedestrians also have high optical flow and hence cannot be filtered out by either method or a combination thereof. Since pedestrians are not labeled in the dataset, they are evaluated as a false positive of the method. However, this is a limitation of the dataset rather than that of the method. The models are class-agnostic by

design i.e. they learn to localize sounds irrespective of the class of objects generating that sound. Attributing corresponding sounds to pedestrians is an expected outcome and it only becomes erroneous due to the way the dataset is annotated.

The performance of optical flow is also seriously hindered by an unstable camera (*Fig.8.b*). A large proportion of videos from the TAU subset of the dataset are taken with a non-stationary camera which makes up the majority of the failure cases of the proposed methods (*Fig.12.b and c, Fig.13*). This is again not necessarily a limitation of the method but rather a characteristic of the dataset that revealed itself while analyzing the shortcomings of the method.

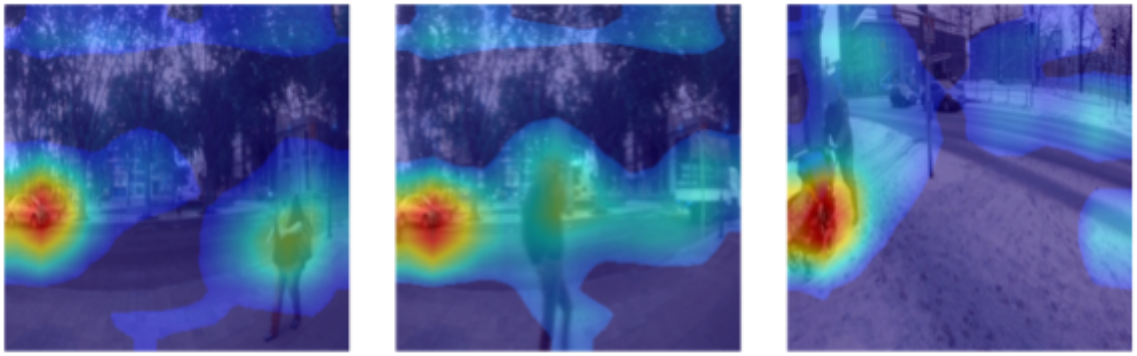

Figure 17. RC-Grad localizing pedestrians as sound sources

The pedestrian case suggests that there is a dissonance between the training and evaluation objectives. The models are trained via contrastive learning to push the representations of sounds and images from the same class close together while driving the same for dissimilar classes farther apart. This is not done for vehicles in particular but for all objects that may appear in the dataset. However, only vehicles are annotated in the dataset, and evaluating against these annotations results in any other sound source, e.g. pedestrians, being a false positive. Also, the models are penalized for these false positives as the metric used to evaluate the localization (IoU) has the area of the union of the prediction and the ground truth in the denominator, and false positives only inflate the denominator resulting in a worse perceived performance of the model.

5.2 Future Work

In this section, potential lines of further research to overcome the limitations addressed in the previous section have been proposed.

5.2.1 Longer-term temporal context

Stationary vehicles with their engines running are a failure case of the proposed method. This is primarily due to the very short term temporal context, 0.125 seconds in our case, that optical flow between consecutive frames provides. Considering that vehicles do not stay stationary for very long periods of time, extending the temporal window to say 5 seconds can give the model the information needed for attributing sounds to vehicles that are temporarily stationary. We consider the following two promising avenues for extending the temporal window.

One simple way is to aggregate optical flow across an expanded window i.e. to calculate optical flow for all pairs of consecutive frames in a window of 5 seconds around the image and use this stack of optical flow as a feature. Similar aggregated flow representations have been used for action recognition [65]. A simpler aggregation strategy is to average optical flow across the time window instead of stacking as done in [41] for estimating object trajectories in the context of sound source localization. A more end-to-end approach would be to ditch optical flow altogether and use a series of frames from the video as an input to the vision encoder making the visual embedding spatio-temporal as done in [39].

5.2.2 Alternate ways of estimating optical flow

We have seen that our results are very sensitive to the quality of optical flow estimation and that optical flow estimation using the Gunnar Farneback algorithm is sensitive to factors like lighting and camera instability. Recent work using deep neural networks for optical flow estimation could be leveraged for improving the quality of optical flow and subsequently the performance of our models. RAFT [62] is the state of the art in optical flow estimation and has been recently added to the pretrained models in PyTorch. Testing the performance of RAFT and other optical flow estimation techniques on Urbansas could yield a performance boost and add robustness to our method.

5.2.3 Evaluating on other datasets

One major limitation of the state-of-the-art methods that we have pointed out is the lack of generalizability to urban scenes due to possible overfitting[4] to common benchmarks. It is only fair to subject our proposed methods to the same scrutiny and evaluate how well they fare in comparison to other methods. We have mentioned that one of our assumptions that sounding objects are also moving objects does not necessarily hold outside of urban scenes. Evaluating on more diverse datasets will help us better understand the strengths and weaknesses of our method and the problem of visual sound source localization in general.

5.2.4 Generating bounding boxes as model predictions

Our models predict arbitrarily shaped masks as localization maps while the ground truth is annotated with bounding boxes. This incongruence is very likely to contribute to a deflation of our results. The arbitrary shapes could be processed to generate bounding boxes through image processing techniques. Also, using our audiovisual embeddings, a supervised model can be trained to predict bounding boxes instead of arbitrary segmentation masks.

[4] The term has been used loosely to suggest that these methods have certain inductive biases that are only applicable to the datasets used to develop them.

Bibliography

[1] Berglund, E., & Sitte, J. (2005, August). Sound source localisation through active audition. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 653-658). IEEE.

[2] Rascon, C., & Meza, I. (2017). Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems*, *96*, 184-210.

[3] Archontis Politis, Kazuki Shimada, Parthasaarathy Sudarsanam, Sharath Adavanne, Daniel Krause, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Yuki Mitsufuji, and Tuomas Virtanen. Starss22: a dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. 2022. URL: https://arxiv.org/abs/2206.01948, arXiv:2206.01948.

[4] Archontis Politis, Annamaria Mesaros, Sharath Adavanne, Toni Heittola, and Tuomas Virtanen. Overview and evaluation of sound event localization and detection in DCASE 2019. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:684–698, 2020. URL: https://ieeexplore.ieee.org/abstract/document/9306885.

[5] Nguyen, T. N. T., Watcharasupat, K. N., Lee, Z. J., Nguyen, N. K., Jones, D. L., & Gan, W. S. (2021). What makes sound event localization and detection difficult? Insights from error analysis. *arXiv preprint arXiv:2107.10469*.

[6] Hershey, J., & Movellan, J. (1999). Audio vision: Using audio-visual synchrony to locate sounds. *Advances in neural information processing systems*, *12*.

[7] Fisher III, John W., et al. "Learning joint statistical models for audio-visual fusion and segregation." *Advances in neural information processing systems* 13 (2000).

[8] Kidron, E., Schechner, Y. Y., & Elad, M. (2005, June). Pixels that sound. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 88-95). IEEE.

[9] Arandjelovic, R., & Zisserman, A. (2017). Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 609-617).

[10] Arandjelovic, R., & Zisserman, A. (2018). Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 435-451).

[11] 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 (pp. 4358-4366).

[12] Oya, T., Iwase, S., Natsume, R., Itazuri, T., Yamaguchi, S., & Morishima, S. (2020). Do we need sound for sound source localization?. In *Proceedings of the Asian Conference on Computer Vision*.

[13] Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., & Zisserman, A. (2021). Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16867-16876).

[14] Mo, S., & Morgado, P. (2022). Localizing Visual Sounds the Easy Way. *arXiv preprint arXiv:2203.09324*.

[15] Wu, Ho-Hsiang, et al. "How to Listen? Rethinking Visual Sound Localization." *arXiv preprint arXiv:2204.05156* (2022).

[16] Fuentes, Magdalena, et al. "Urban sound & sight: Dataset and benchmark for audio-visual urban scene understanding." *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

[17] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012).

[18] Hannun, A. et al. Deep speech: scaling up end-to-end speech recognition. Preprint at arXiv https://arxiv.org/abs/1412.5567 (2014).

[19] Radford, A. et al. Better language models and their implications. OpenAI Blog https://openai.com/blog/better-language-models/ (2019).

[20] Finn, C., Goodfellow, I. & Levine, S. Unsupervised learning for physical interaction through video prediction. Adv. Neural Inf. Proc. Sys. 29, 64–72 (2016).

[21] Silver, D. et al. Mastering the game of Go without human knowledge. Nature 550, 354–359 (2017).

[22] Santoro, A. et al. A simple neural network module for relational reasoning. Adv. Neural Inf. Proc. Sys. 30, 4967–4976 (2017).

[23] Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798-1828.

[24] Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*, *117*(48), 30033-30038.

[25] Marr, D. (1982). *Vision. A computational investigation into the human representation and processing of visual information*. San Francisco: Freeman.

[26] Geirhos, R., et al. "Shortcut learning in deep neural networks." *Nature Machine Intelligence* 2.11 (2020): 665-673.

[27] Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision* (pp. 843-852).

[28] Ericsson, L., Gouk, H., Loy, C. C., & Hospedales, T. M. (2022). Self-Supervised Representation Learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, *39*(3), 42-62.

[29] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[30] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

[31] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

[32] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2536-2544).

[33] Gidaris, S., Singh, P., & Komodakis, N. (2018, February). Unsupervised Representation Learning by Predicting Image Rotations. In *International Conference on Learning Representations*.

[34] Dosovitskiy, A., Springenberg, J. T., Riedmiller, M., & Brox, T. (2014). Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, *27*.

[35] Song, Z., Wang, Y., Fan, J., Tan, T., & Zhang, Z. (2022). Self-Supervised Predictive Learning: A Negative-Free Method for Sound Source Localization in Visual Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3222-3231).

[36] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1993). Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, *6*.

[37] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). Ieee.

[38] Qian, R., Hu, D., Dinkel, H., Wu, M., Xu, N., & Lin, W. (2020, August). Multiple sound sources localization from coarse to fine. In *European Conference on Computer Vision* (pp. 292-308). Springer, Cham.

[39] Senocak, A., Ryu, H., Kim, J., & Kweon, I. S. (2022). Less Can Be More: Sound Source Localization With a Classification Model. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 3308-3317).

[40] Korbar, B., Tran, D., & Torresani, L. (2018). Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, *31*.

[41] Afouras, T., Owens, A., Chung, J. S., & Zisserman, A. (2020, August). Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision* (pp. 208-224). Springer, Cham.

[42] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and An- tonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In Proc. ICCV, 2019

[43] Valverde, Francisco Rivera, Juana Valeria Hurtado, and Abhinav Valada. "There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[44] Zürn, Jannik, and Wolfram Burgard. "Self-Supervised Moving Vehicle Detection from Audio-Visual Cues." arXiv preprint arXiv:2201.12771 (2022).

[45] Zinemanas, P., Cancela, P., & Rocamora, M. (2019). MAVD: A dataset for sound event detection in urban environments. *Detection and Classification of Acoustic Scenes and Events, DCASE 2019, New York, NY, USA, 25–26 oct, page 263--267*.

[46] Wang, S., Mesaros, A., Heittola, T., & Virtanen, T. (2021, June). A curated dataset of urban scenes for audio-visual scene analysis. In *ICASSP 2021-2021*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 626-630). IEEE.

[47] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 658-666).

[48] Wang, C. Y., Yeh, I. H., & Liao, H. Y. M. (2021). You only learn one representation: Unified network for multiple tasks. *arXiv preprint arXiv:2105.04206*.

[49] Chen, H., Xie, W., Vedaldi, A., & Zisserman, A. (2020, May). Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 721-725). IEEE.

[50] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

[51] Shah, S. T. H., & Xuezhi, X. (2021). Traditional and modern strategies for optical flow: an investigation. *SN Applied Sciences*, *3*(3), 1-14.

[52] Gibson, J. J., & Carmichael, L. (1966). *The senses considered as perceptual systems* (Vol. 2, No. 1, pp. 44-73). Boston: Houghton Mifflin.

[53] Horn, B. K., & Schunck, B. G. (1981). Determining optical flow. *Artificial intelligence*, *17*(1-3), 185-203.

[54] Horn, B., Klaus, B., & Horn, P. (1986). *Robot vision*. MIT press.

[55] B. Lucas, T. Kanade, An iterative technique of image registration and its application to stereo (1981). The 7th International Joint Conference on Artificial Intelligence, IJCAI. (1981) 674–679.

[56] T. Brox, A. Bruhn, N. Papenberg, J. Weickert, High accuracy optical flow estimation based on a theory for warping, volume 3024, 2004, pp. 25–36

[57] C. Bailer, B. Taetz, D. Stricker Flow fields: dense correspondence fields for highly accurate large displacement optical flow estimation IEEE Trans. Pattern Anal. Mach. Intell., 41 (8) (2019), pp. 1879-1892

[58] Farnebäck, Gunnar. "Two-frame motion estimation based on polynomial expansion." Scandinavian conference on Image analysis. Springer, Berlin, Heidelberg, 2003.

[59] Dosovitskiy, Alexey, et al. "Flownet: Learning optical flow with convolutional networks." *Proceedings of the IEEE international conference on computer vision*. 2015.

[60] Ilg, Eddy, et al. "Flownet 2.0: Evolution of optical flow estimation with deep networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[61] Sun, Deqing, et al. "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[62] Teed, Zachary, and Jia Deng. "Raft: Recurrent all-pairs field transforms for optical flow." *European conference on computer vision*. Springer, Cham, 2020.

[63] Kajo, I., Malik, A. S., & Kamel, N. (2016, August). An evaluation of optical flow algorithms for crowd analytics in surveillance system. In *2016 6th International conference on intelligent and advanced systems (ICIAS)* (pp. 1-6). IEEE.

[64] Xiao, F., & Jae Lee, Y. (2016). Track and segment: An iterative unsupervised approach for video object proposals. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 933-942).

[65] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, *27*.

[66] Sun, S., Kuang, Z., Sheng, L., Ouyang, W., & Zhang, W. (2018). Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1390-1399).

[67] Liu, C., & Sun, D. (2011, June). A bayesian approach to adaptive video super resolution. In *CVPR 2011* (pp. 209-216). IEEE.

[68] Amerini, I., Galteri, L., Caldelli, R., & Del Bimbo, A. (2019). Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF international conference on computer vision workshops* (pp. 0-0).

[69] Lamdouar, H., Yang, C., Xie, W., & Zisserman, A. (2020). Betrayed by motion: Camouflaged object discovery via motion segmentation. In *Proceedings of the Asian Conference on Computer Vision*.

[70] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740-755). Springer, Cham.