Self-View Grounding Given
a Narrated 360 Degree Video

Shih-Han Chou,[†] **Yi-Chun Chen,**[†] **Kuo-Hao Zeng,**[†]
Hou-Ning Hu,[†] **Jianlong Fu,**[‡] **Min Sun**[†]

[†]Department of Electrical Engineering, National Tsing Hua University
[‡]Microsoft Research, Beijing, China
{happy810705, yichun8447}@gmail.com, khzeng@cs.stanford.edu
{eborboihuc@gapp, sunmin@ee}.nthu.edu.tw, jianf@microsoft.com

Abstract

Narrated 360° videos are typically provided in many touring scenarios to mimic real-world experience. However, previous work has shown that smart assistance (i.e., providing visual guidance) can significantly help users to follow the Normal Field of View (NFoV) corresponding to the narrative. In this project, we aim at automatically grounding the NFoVs of a 360° video given subtitles of the narrative (referred to as "NFoV-grounding"). We propose a novel Visual Grounding Model (VGM) to implicitly and efficiently predict the NFoVs given the video content and subtitles. Specifically, at each frame, we efficiently encode the panorama into feature map of candidate NFoVs using a Convolutional Neural Network (CNN) and the subtitles to the same hidden space using an RNN with Gated Recurrent Units (GRU). Then, we apply soft-attention on candidate NFoVs to trigger sentence decoder aiming to minimize the reconstruct loss between the generated and given sentence. Finally, we obtain the NFoV as the candidate NFoV with the maximum attention without any human supervision. To train VGM more robustly, we also generate a reverse sentence conditioning on one minus the soft-attention such that the attention focuses on candidate NFoVs less relevant to the given sentence. The negative log reconstruction loss of the reverse sentence (referred to as "irrelevant loss") is jointly minimized to encourage the reverse sentence to be different from the given sentence. To evaluate our method, we collect the first narrated 360° videos dataset and achieve state-of-the-art NFoV-grounding performance.

Introduction

Thanks to the availability of consumer-level 360° video cameras, many 360° videos are shared on websites like YouTube and Facebook. Among these videos, a subset of them is narrated with natural language phrases describing the video content. For instance, many videos consist of the guided tour of real estates (e.g., houses and apartments) or tourist locations. Intuitively, a 360° touring video provides viewers the freedom to follow the narrative and select Normal Field of Views (NFoVs) provided by typical video players (see Fig. 1). However, a study in Lin et al. (2017) suggests that visual guidance (i.e., assistance through visual indicator) is preferred to assist viewers to follow the NFoV described by the narrative. In order to provide the visual
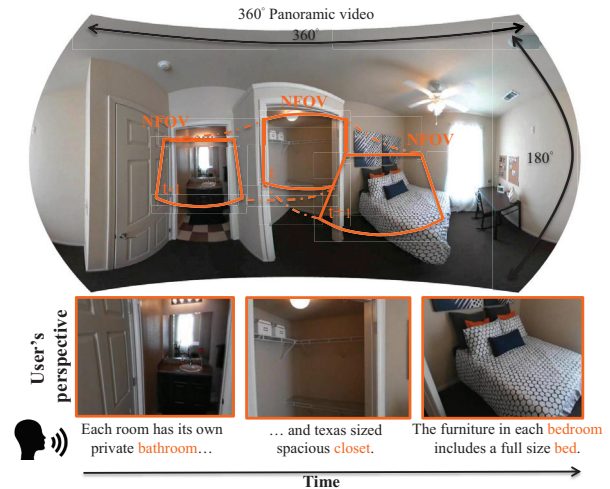
Figure 1: **Illustration of NFoV-grounding**. In a 360° video (top-panel), a video player displays a predefined Normal Field of View (NFoV) (bottom-panel). Our VGM can automatically ground narrative into a corresponding NFoV (orange boxes) at each frame.

guidance, the NFoV described by the narrative needs to be inferred. We define this new task as "NFoV-grounding".

The task of NFoV-grounding is related to but different from grounding language in normal images in two main ways. First of all, the panoramic image in 360° video has a large field of view with high resolution. Hence, the regions described by narrative often are relatively small compared to the panoramic image. As a result, grounding language in panoramic image is harder and computationally expensive. Secondly, rather than grounding to an object, our task is grounding to an NFoV, which could correspond to objects and/or scene regions. In this case, existing object proposal methods cannot be leveraged. To the best of our knowledge, NFoV-grounding is a unique task which hasn't been tackled.

We propose a novel Visual Grounding Model (VGM) to implicitly and efficiently predict the NFoVs given the video content and subtitles. Specifically, at each frame, we encode the panorama into feature map using a Convolutional Neural Network (CNN) and embed candidate NFoVs onto the

feature map to efficiently process NFoVs in parallel. On the other hand, the subtitle is encoded to the same hidden space using an RNN with Gated Recurrent Units (GRU). Then, we apply soft-attention on candidate NFoVs to trigger a sentence decoder aiming to minimize the reconstruction loss between the generated and given sentence (referred to as relevant loss). At the end, we obtain the NFoV as the candidate NFoV with the maximum attention without any human supervision. We emphasize that the training does not require the knowledge of ground truth NFoV. Similar to Rohrbach et al. (2016), our model relies on sentence reconstruction to implicitly infer the NFoV. In order to further address the challenges in NFoV-grounding, we propose the following techniques to train VGM more robustly. First of all, we generate a "reverse sentence" conditioning on one minus the soft-attention such that the attention focuses on candidate NFoVs less relevant to the given sentence. The negative log reconstruction loss of the reverse sentence (referred to as "irrelevant loss") is jointly minimized to encourage reverse sentences to be different from the given sentences. Secondly, we augment the panoramic images dataset by exploiting rotation invariant property to randomly shift the viewing angles.

To evaluate our method, we collect the first narrated 360° video dataset consisting of both indoor and outdoor tourist guides. We also redefine recall and precision on NFoV-grounding as evaluation metrics. Finally, our model achieves state-of-the-art NFoV-grounding performance and can be run at 0.38 fps on 720×1280 panoramic image.

Our main contributions can be summarized as follows:

- We define a new "NFoV-grounding" task which is essential to automatic assist watching 360° videos.

- We propose a novel Visual Grounding Model (VGM) to implicitly and efficiently infer NFoV.

- We introduce a novel irrelevant loss and a 360° data augmentation technique to robustly train our model.

- We collect the first narrated 360° video dataset and achieve the best performance.

Related Work

We review related works in virtual cinematography, 360° vision and grounding natural language in images and videos.

Virtual Cinematography

A listing of virtual cinematography research (Christianson et al. 1996; Elson and Riedl 2007; Mindek et al. 2015) focused on controlling a camera view in virtual/gaming environments and did not take the problem of perception difficulties into consideration. Some works (Sun et al. 2005; Chen and Carr 2015; Chen et al. 2016) relaxed such perception assumption. They manipulated virtual cameras within a static video with wide field-of-view of a teleconference, a basketball court, or a classroom, where objects of interest could be easily extracted.

360° Vision

Recently, the perception/experience of 360° vision is gained numerous interest. Assens et al. (Assens et al. 2017) focused on scan-paths prediction on 360° images. The network predicts saliency volumes, which are stacks of saliency maps, then the scan-paths were sampled from the volume conditioned on number, location, and duration of fixations. Lin et al. (Lin et al. 2017) concluded that the use of Focus Assistance for 360° videos will help viewers to focus on the intended targets in videos. Su et al. (Su, Jayaraman, and Grauman 2016) referred a problem of viewing NFoV in 360° videos as Pano2Vid and proposed an offline method handling unedited 360° videos downloaded from YouTube. They further (Su and Grauman 2017) improved the offline method in threefold. First, they proposed a coarse-to-fine technique to reduce computational costs. Second, they gave diverse output trajectories. Last, they introduced a degree of freedom by zooming NFoV. In contrast, Hu and Lin et al. (Hu et al. 2017) proposed an online human-like agent piloting through 360° videos. They argued that a human-like online agent is necessary in order to provide more effective video-watching supports for streaming videos and other human-in-the-loop applications, such as foveated rendering (Patney et al. 2016). On the other hand, Lai et al. (Lai et al. 2017) provided an offline editing tool on 360° videos, which equips with visual saliency and semantic scene label, helping end users generate a stabilized NFoV video hyperlapse while our method focuses on semantic automatic visual NFoV-grounding.

Grounding natural language in images and videos

The interaction between natural language and vision has been extensively studied over the past years (Zeng et al. 2016; 2017). For grounding language in images, (Johnson et al. 2015) used a Conditional Random Field model to ground a scene graph query of images in order to retrieve semantically related images. (Karpathy, Joulin, and Li 2014) reasoned dependency tree relations on images using Multiple Instance Learning and a ranking objective. (Karpathy and Li 2015) replaced the dependency tree with a multi-modal Recurrent Neural Network and simply used maximum value instead of ranking objective. (Wang, Li, and Lazebnik 2016) proposed a structure-preserving image-sentence embedding method for retrieval problem and also applied it to phrase localization. (Mao et al. 2016) and the Spatial Context Recurrent ConvNet (SCRC) (Hu et al. 2016) used a recurrent caption generation network to localize an object with the highest probability by evaluating the given phrase on the set of proposal boxes. (Rohrbach et al. 2016) proposed an attention localization mechanism with an extra text reconstruction task. (Rong, Yi, and Tian 2017) proposed models for both scene text localization and retrieving candidate text regions by jointly scoring and ranking the text outputs. (Zhang et al. 2017) associated image regions with text queries by a discriminative bimodal neural network, with extensive use of negative samples. (Xiao, Sigal, and Jae Lee 2017) localized textual phrases in a weakly-supervised setting by learning pixel-level spatial attention masks as phrases localization. Several representative works on spatial-temporal language grounding are (Lin et al. 2014a), (Yu and Siskind 2013), and (Li et al. 2017) aimed at tracking objects of interest from a sequence of natural language specification while

ours focus on visual NFoV-grounding where might deal with indoor or outdoor scenery and multiple challenging NFoVs of interest.

Method

Our goal is to build a system that can ground subtitles (i.e., natural language phrases) into corresponding views in each 360° video. We define this task as the Normal Field of View (NFoV)-grounding problem since typical 360° video players display a predefined NFoV (see Fig. 1). We formally define the notation and task below.

Notation. We define $\mathbf{p}$ as a sequence of panoramic frames where $\mathbf{p} = \{p^1, p^2, ..., p^k\}_{k=1}^{K}$ and K is the video length. $\mathbf{f} = \{f^1, f^2, ..., f^k\}$ is defined as the encoded panoramic visual feature map. $\mathbf{v}^k = \{v_1^k, v_2^k, ..., v_i^k\}$ is defined as the encoded NFoV candidates' visual feature. $\mathbf{S}$ is the subtitles where $\mathbf{S} = \{\mathbf{s}^1, \mathbf{s}^2, ..., \mathbf{s}^k\}$, $\mathbf{s} = \{s_1, s_2, ..., s_m\}$, and m is the number of words in a subtitle. $\mathbf{L} = \{\mathbf{l}^1, \mathbf{l}^2, ..., \mathbf{l}^k\}$ is defined as the encoded language features, where $\mathbf{l} = \{l_1, l_2, ..., l_m\}$. $\boldsymbol{\alpha}$ is defined as the soft-attention weights where $\boldsymbol{\alpha}^k = \{\alpha_1^k, \alpha_2^k, ..., \alpha_i^k\}$. v_{att} is the attended NFoV's visual feature and $\hat{v}_{att}^k$ is the reverse attended NFoV's visual feature. v_{rec}^k is the reconstructed NFoV's visual feature and $\hat{v}_{rec}$ is the reverse reconstructed NFoV's visual feature. $\mathbf{y}$ is the predicted NFoVs. A panoramic frame usually has a corresponding subtitle describing several objects/regions and the relation between them.

Task: NFoV-grounding. Given panoramic frames and subtitles, our work is to ground the NFoV viewpoint based on the subtitles in panoramic frames to provide visual guidance:

$$\mathbf{y} = O(\mathbf{p}, \mathbf{S}), \tag{1}$$

where O denotes a model inputting panoramic frames $\mathbf{p}$ and subtitles $\mathbf{S}$ and predicting the corresponding NFoVs $\mathbf{y}$.

Model Overview

We now present an overview of our model. Our model consists of three main components. The first part encodes each panoramic frame and its corresponding subtitle into a hidden space with the same dimension. For visual encoding, every panoramic frame p^k is encoded to a panoramic visual feature map f^k by a convolutional neural network (CNN); for language encoding, the corresponding subtitle $\mathbf{s}^k$ is encoded to language feature l^k by a recurrent neural network (RNN). After encoding, we have the panoramic visual feature map f^k and the subtitle representation l^k (see Fig. 2(a)).

The second part is the proposed Visual Grounding Model (VGM) (see Fig. 2(b)). Similar to (Su, Jayaraman, and Grauman 2016), we define several NFoV candidates centering at longitudes $\phi \in \boldsymbol{\phi} = \{0, 30, 60, ..., 330\}$ and latitudes $\theta \in \boldsymbol{\theta} = \{0, \pm 15, \pm 30\}$. Then, we propose to encode NFoV visual feature candidate $\mathbf{v}^k$ from panoramic visual feature map f^k. Note that each NFoV corresponds to a rectangle region on an image with a perspective projection, but a distorted region on the panoramic image with equirectangular projection. Hence, we propose to embed an NFoV generator proposed in (Su, Jayaraman, and Grauman 2016) into representation space to efficiently encode the panoramic visual

feature map f^k into NFoV candidates' visual feature $\mathbf{v}^k$:

$$\mathbf{v}^k = G_{NFoV}(f^k). \tag{2}$$

Once we have NFoV candidates' visual feature $\mathbf{v}^k$, the VGM applies the soft-attention mechanism to the encoded NFoV candidates' $\mathbf{v}^k$ guided by the encoded subtitle language features $\mathbf{l}^k$ to obtain the attended weights $\boldsymbol{\alpha}^k$.

The third part is to reconstruct the subtitles. At each frame k, we reconstruct the corresponding subtitles by inputting $\boldsymbol{\alpha}^k$ to language decoder during the training phase. Using the attended feature $\boldsymbol{\alpha}^k$ derived from the VGM, we further reconstruct the subtitle. After the reconstruction, we compute the similarity between the original input subtitles and the reconstructed ones. In this case, our goal is to maximize the similarity between them so that we can learn a model specializing in NFoV grounding without direct supervision. On the other hand, during testing phase, we acquire the selected NFoV among $\mathbf{v}^k$ by attention scores during testing phase.

Next, we describe each component in details.

Panoramic/Subtitle Encoder

The goal of this part is to encode the panoramic frame and the subtitle into the same hidden space. Given a 360° video panoramic frame and its corresponding subtitles, we use the CNN to encode the 360° frame and use the RNN to encode the subtitles. Every panoramic frame is then encoded into a panoramic visual feature map f:

$$f^k = CNN(p^k). \tag{3}$$

We utilize ResNet-101 (He et al. 2016) as visual encoder. For subtitle, we first extract word representation by a pretrained word embedding model(Pennington, Socher, and Manning 2014). Then we encode the subtitle's representation in turn with an Encoder-RNN (RNN_E) and obtain the subtitle representation. The language feature l^k is as follow:

$$\mathbf{l}^k = RNN_E(\mathbf{s}^k). \tag{4}$$

In practice, we employ Gated Recurrent Units (GRU) (Chung et al. 2014) as our language encoder.

Visual Grounding Model

This part is the proposed Visual Grounding Model (VGM). The goal of this part is to ground the NFoV in the panorama with the corresponding subtitles. After deriving the encoded panoramic visual feature map f^k and the language feature $\mathbf{l}^k$, we use an NFoV generator to generate the encoded NFoV candidates' visual feature $\mathbf{v}^k$ directly from feature map. The original function of the NFoV generator is to retrieve the pixels in the panoramic frame from a given viewpoint. Proposing NFoV at pixel space and training a visual grounding model like (Rohrbach et al. 2016) would impose three drawbacks: (1) large memory requirement for the model, (2) considerable time consuming for both training and testing, and (3) making end-to-end training infeasible. As a result, we embed the NFoV generator from pixel space into feature space to mitigate those issues. We propose 60 spatial glimpses, which are at longitudes ϕ and latitudes $\boldsymbol{\theta}$ directly
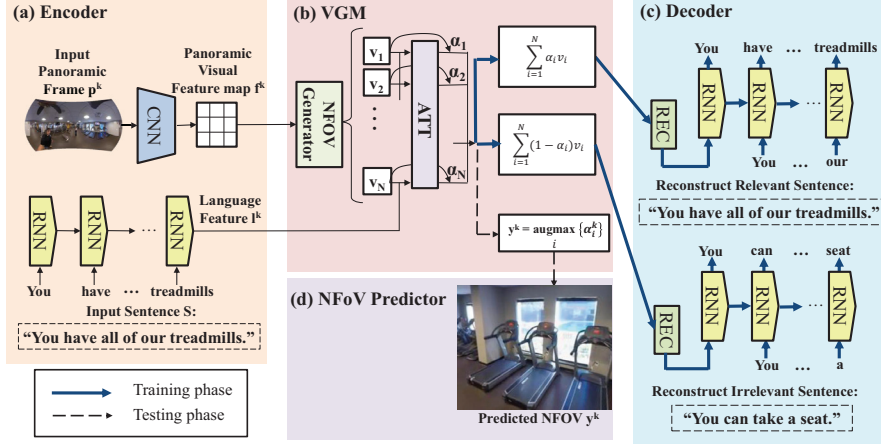
Figure 2: Illustration of our method. During unsupervised training, Each 360 video frame pass through the panel (a) panoramic/subtitle encoder, in which given panoramic frames are encoded to visual feature while corresponding subtitles are encoded to language feature by CNN and RNN respectively, (b) our proposed VGM, in which we propose NFoV candidates and apply the soft-attention mechanism on NFoV candidates with encoded language feature, and (c) language decoder, in which we reconstruct subtitles according to attended feature derived from VGM. In the testing phase, instead of language decoder, (d) NFoV predictor is used.

on feature map. Once we have these NFoV candidates' visual feature $\mathbf{v^k}$, we calculate the attention of every NFoV candidate by the corresponding visual feature v_i^k and the subtitle representation $\mathbf{l}^k$. Then, we use the two layers perceptron to compute the attention of each NFoV candidate:

$$\bar{\alpha}_i^k = ATT(v_i^k, \mathbf{l}^k) = W_a\sigma(W_v v_i^k + W_l \mathbf{l}^k + b_1) + b_a, \quad (5)$$

where σ is the hyperbolic function, and W_v, W_l are the parameters of two fully connected layers. Then we employ softmax to get the normalized attention weights:

$$\alpha_i^k = softmax(\bar{\alpha}_i^k). \quad (6)$$

After attaining the attention distribution, we compute attention feature v_{att}^t by calculating the weighted sum of the visual features v^t and the attention weights α^t.

$$v_{att}^k = \sum_{i=1}^{N} \alpha_i^k v_i^k, \quad (7)$$

where N is the number of NFoV candidates. Furthermore, instead of only reconstructing the subtitles from the most relevant visual feature, we also generate a reverse sentence conditioning on one minus the attention scores such that the attention focuses on irrelevant candidate NFoVs:

$$\hat{\alpha}_i^k = 1 - \alpha_i^k. \quad (8)$$

Then we also calculate the weighted sum to compute the reverse attention feature $\hat{v}_{att}^k = \sum_{i=1}^{N} \hat{\alpha}_i^k v_i^k$.

During testing phase, our model predicts an NFoV y from all NFoV candidates. Because the correct NFoV prediction must contribute the most visual information to reconstruct the corresponding subtitles, our model predicts y by selecting the NFoV having the highest attention score:

$$y^k = \underset{i}{\mathrm{argmax}} \ \{\alpha_i^k\}. \quad (9)$$

Reconstruct Subtitles

As demonstrated in (Rohrbach et al. 2016), learning to reconstruct the descriptions of objects within an image shows impressive results on visual grounding task. As a result, we mimic them to employ the reconstruction loss to perform unsupervised learning on NFoV grounding task in 360° video.

First, we obtain the reconstruct feature v_{rec}^k and the reverse reconstruct feature $\hat{v}_{rec}^k$ after encoding the attention feature by a non-linear encoding layer:

$$v_{rec}^k = \sigma(W_r v_{att}^k + b_r); \hat{v}_{rec}^k = \sigma(W_r \hat{v}_{att}^k + b_r). \quad (10)$$

We then employ a Decoder-RNN as our language decoder (RNN_D). The output dimension of such decoder is set as the size of the dictionary. Thus, our language decoder takes reconstructed feature v_{rec}^k and reverse reconstructed feature $\hat{v}_{rec}^k$ as inputs to generate a distribution over subtitle s_t:

$$P(\mathbf{s}_k|v_{rec}^k) = RNN_D(v_{rec}^k); P(\mathbf{s}_k|\hat{v}_{rec}^k) = RNN_D(\hat{v}_{rec}^k), \quad (11)$$

where $P(\mathbf{s}_k|v_{rec}^k)$ and $P(\mathbf{s}_k|\hat{v}_{rec}^k)$ are distributions over the words conditioned on the input reconstructed feature and reverse reconstructed feature, respectively. In practice, we utilize Long Short-Term Memory Units (LSTM) (Hochreiter and Schmidhuber 1997) as our language decoder. Thanks to lots of research on image captioning (Vinyals et al. 2015; Xu et al. 2015) having demonstrated the effectiveness of LSTM on image captioning task, we first pre-train our language decoder on an image captioning dataset. The pre-trained decoder provides us faster training and better performance on NFoV-grounding task.

Finally, we train our network by maximizing/minimizing the likelihood of the corresponding subtitle $\mathbf{s}_k$ generated via input feature $v_{rec}^k/\hat{v}_{rec}^k$ during reconstruction. In this case,

Indoor Subset			Outdoor Subset		
Here we have multiple computers.	Also we have nice printing, a nice drop-down projection screen,	and plenty of table space for you to come and study and relax.	One of the best vantage points to admire the Eiffel Tower.	You can see the Palais Chaillot with: on the left the Museum of Man,	and right the City of Architecture and Heritage.
Over on this side you will see one of our really, really cool, healthy vending machines.	The TV as well as a really, really nice long sectional for everybody.	There is additional counter space.	If you look around you will see our pool to say it is a good size.	The deck area has lots of area where you can enjoy and relax sunbathe.	You have your grilling area where you can grill and your fire pit

Figure 3: Annotation examples. In the left panel shows indoor videos, while outdoor videos are shown in the right panel. Human annotators are asked to annotate the objects or places they would like to see when given the narratives.

the overall loss function can be defined as:

$$L = \frac{1}{B}\sum_{b=1}^{B}[-\lambda P(\mathbf{s}_k|v_{rec}^k)+(1-\lambda)log(P(\mathbf{s}_k|\hat{v}_{rec}^k))], \quad (12)$$

where B and λ denote batch size and the controlled ratio balancing the effect between relevant and irrelevant visual feature. Because the relevant loss has a lower bound ($zero$), it will converge to the lower bound when we minimize it. However, the irrelevant loss is unbounded ($-infinite$), so when we try to minimize both losses, the irrelevant loss would dominate. Therefore, we mitigate this issue by adding a log function to the irrelevant loss.

Training with Data Augmentation on $360°$ image

Since $360°$ images are panoramic, we are allowed to arbitrarily augment training data by rotating an entire panoramic image along the longitude. In practice, we online randomly select a value $x \in [0, X]$ where X is the length of a panoramic image. Then, we paste the pixels on the left-hand side of x to the end of the right-hand side. This operation performs the rotating along the longitude centering on x. Since we do this operation online, no data is stored in advance. As a result, our augmentation method on $360°$ frames provides memory-free data augmentation.

Implementation Details

We train our model only relying on Eq. (12) which does not contain any supervision for NFoV grounding. We set $\lambda = 0.8$. This is set empirically without heavily tuning. We decrease the frame rate to 1 to save memory usage and set dictionary dimension as 9956 according to the number of words appearing in all subtitles. We randomly sample 3 consecutive frames during training phase (i.e., $k = 3$), but evaluate all frames during testing phase (i.e., $k = \#frames$). Since the maximal length of subtitles is 33, we set $m = 33$ and give the remaining empty words Pad token if the length of subtitle less than 33. Besides, we add $Start$ and End tokens to represent the beginning and end of a subtitle, respectively. We use Adam (Kingma and Ba 2015) as opti-

mizer with default hyperparameters and 0.001 learning rate and set batch size B by 4. We use ResNet-101 pre-training on ImageNet (Deng et al. 2009) as our visual encoder and we pre-train our language decoder on MSCOCO dataset (Lin et al. 2014b). We implement all of our methods by Py-Torch (Paszke and Chintala). The training phase costs about $17\ min$ and $7\ sec$ per epoch and please refer to the Sec. Modal Efficiency for further model's efficiency analysis.

Dataset

In order to evaluate our method, we collect the first narrated $360°$ videos dataset. This dataset consists of touring videos, including scenic spots and housing introduction, and subtitles files, including subtitle text and start and end timecode. Both the videos and the subtitle files are downloaded from Youtube, noted that some of the subtitles are created by Youtube's speech recognition technology.

The videos are separated into two categories: indoor and outdoor, and resized to 720×1280 first. We extracted a continuous video clip from each video where a scene transition is absent. Subtitle files are also clipped into several files according to the transition time. For training data, we select those whose duration is within 90% of the range of the duration of all videos, so the max video length is 44 seconds. Also, since the outdoor touring videos are easy to contain some uncommon words or non-English words, our model is hard to learn the word to find the best viewpoint. Hence, we use WordNet as the criterion for outdoor training videos, only sampling the videos without uncommon words (words not in WordNet). For validation and testing data, both video segments and their annotated ground truth objects are included. We ask human annotators to annotate the objects mentioned in narratives on panoramas chronologically according to the start and end time code in subtitle files. Example panoramas and subtitles are shown in Fig. 3. Finally, we have 563 indoor videos and 301 outdoor videos. We assign 80% of the videos and subtitles for training and 10% each for validation and testing. (Available at http://aliensunmin.github.io/project/360grounding/)

Table 1: Our 360 Videos Dataset: we list the statistics information of our dataset. We separate train/val/test set, and they all contain indoor/outdoor set. In the ground truth annotation, we do not label the training set thus the number of it is unknown.

	Train		Validation		Test	
	Indoor	Outdoor	Indoor	Outdoor	Indoor	Outdoor
Video length (in average)	16.4	13.87	18.1	15.9	23.4	22.2
Video length (maximum)	44	44	37	35	83	76
subtitle length (in average)	11.6	10.64	11.4	9.67	10.5	10.9
subtitle length (maximum)	33	32	38	29	30	51
#Ground truth annotation (in average)	–	–	2.95	1.49	3.18	1.64
#Videos	466	216	41	41	56	44

Table 2: Ablation Studies. We evaluate several variants of the proposed model. The w/ denotes with.

Method / Model	RL	RL-f	$D^{\dagger}$-RL-f	$D^{\dagger}$-RIL-f
w/ Relevant Loss	√	√	√	√
w/ Embedded NFoV generator	–	√	√	√
w/ Data augmentation	–	–	√	√
w/ Irrelevant Loss	–	–	–	√

Experiments

Because the style of indoor videos and outdoor videos are different on both vision and subtitles, we first conduct the ablation studies of our proposed method and compare our model with baselines in the beginning. Then, we compare our best model with baselines on the total dataset (i.e., the combination of indoor and outdoor subsets) to demonstrate the robustness of our proposed method. In the end, we manifest the efficiency of our proposed method by measuring the speed of our best model. In the following, we first describe the baseline methods and variants of our method. Then, we define the evaluation metrics. Finally, we show the results and make a brief discussion.

Baselines

- *RS:* Random Selection. We randomly select the NFoV candidates for fundamental evaluation.
- *CS:* Centric Selection. We evaluate the dataset by selecting centric NFoV candidate as predicted NFoV. The effectiveness of centric selection in many visual tasks has been demonstrated by lots of literature (Li, Fathi, and Rehg 2013; Judd et al. 2009).
- *RL:* Relevant Loss (Rohrbach et al. 2016). We implement the model proposed by (Rohrbach et al. 2016) as a baseline. We replace region proposals by NFoV proposal and follow the same experimental setting for fair comparison.

Ablation studies

We also evaluate several variants of the proposed model. Note that *w/* denotes with. The variants of models are listed in Tab. 2 and the details are as follows:
- *w/ Relevant Loss:* Train the model with the relevant loss proposed in (Rohrbach et al. 2016).
- *w/ Embedded NFoV generator:* Embed the NFoV generator into feature space (See Sec. Visual Grounding Model).
- *w/ Irrelevant Loss:* Train the model with the irrelevant loss (See Eq. (11)).

Table 3: Average Recall and Precision in Indoor subset. Our model with data augmentation, pretrained decoder, relevant/irrelevant loss and NFoV generator embedding achieves the highest recall/precision over all baselines and proposed methods.

Model	avg. Recall (%)	avg. Precision (%)
RS	8.8	21.5
CS	10.3	29.5
RL	6.1	12.6
RL-f	8.3	19.2
$D^{\dagger}$-RL-f	12.8	27.8
$D^{\dagger}$-RIL-f	13.4	30.7
Oracle	51.1	70.0

- *w/ Data augmentation:* Train the model with augmented training data.

Evaluation Metrics

In this work, we are interested in predicting the objects or places mentioned in subtitles. Because the objects or places in the proposed dataset are annotated by bounding boxes, they typically are not contained in an NFoV, even appear in multiple NFoVs. In this case, we propose to utilize recall and precision measurement calculated at the pixel level to evaluate the performance of our model and baselines. The recall and precision can be defined as:

$$Recall = \frac{\mathbf{GT_{bbox}} \cap y}{\mathbf{GT_{bbox}}}; Precision = \frac{\mathbf{GT_{bbox}} \cap y}{y}, \tag{13}$$

where y denotes the predicted NFoV (See Eq. (9)), $\mathbf{GT_{bbox}}$ denotes the grounding bounding boxes annotated by human annotator, and $\mathbf{GT_{bbox}} \cap y$ means compute the overlap between annotated bounding boxes and predicted NFoV by pixels. We compute $Recall$ and $Precision$ each frame and average them across all frames in a video to acquire $avg.Recall$ and $avg.Precision$. Besides, to quantify the speed of our proposed model, we also measure fps in the testing set. We conduct all experiments on a single computer with Intel(R) Core(TM) i7-5820K CPU @ 3.30GHz, 64GB RAM DDR3, and an NVIDIA TitanX GPU.

Results and Discussion

Ablation studies. The results are shown in Tab. 3 verify that our proposed method can better ground language in

Table 4: Average Recall and Precision on the total dataset. We test our model in total dataset (indoor + outdoor) and our model outperforms the baselines on both $avg.Recall$ and $avg.Precision$.

Model	avg. Recall (%)	avg. Precision (%)
RS	8.7	20.6
CS	9.7	28.2
RL	5	12.9
D†-RIL-f	18.1	34.6
Oracle	51.7	69.4

Table 5: We record the testing time over total dataset and compare our proposed model with the strongest baseline. We resize the image to the different ratio and report their fps and grounding performance. The full size denotes 720×1280 panoramic image. Evidently, our proposed model is more efficient than the strongest baseline. In $1/16$ image size, ours fps is 10 times faster than baseline but has the same performance.

RL	fps	avg. Recall / avg. Precision (%)
Full size	infeasible	infeasible
1/4	infeasible	infeasible
1/16	0.14	5/12.9
D†-RIL-f	fps	avg. Recall / avg. Precision (%)
Full size	0.38	18.1 / 34.6
1/4	1.11	6.3 / 20
1/16	1.4	5.4/15.9

the panoramic image than typical visual grounding method. Moreover, the ablation studies demonstrate that all of our proposed techniques improve the performance. Our best model (D†-RIL-f) outperforms the strongest baseline (CS) by 30% gain on $avg.Recall$.

Model Robustness. Tab. 4 shows that our best model (D†-RIL-f) achieves the state-of-the-art performance (18.1% and 34.6% on $avg.Recall$ and $avg.Precision$) on the first narrated $360°$ video dataset. Our best model (D†-RIL-f) is significantly better than the strongest baseline (CS) by 8.4% and 6.4% on $avg.Recall$ and $avg.Precision$, respectively. It manifests the robustness of our proposed methods.

Model Efficiency. The results are shown in Tab. 5 illustrate that our model is more effective and efficient than baseline. The baseline needs too many memories to train and evaluate on a single computer at the full size and $1/4$ ratio setting. Besides, our best model (D†-RIL-f) achieves 0.38 fps on the full size panoramic image (720×1280). Also, our model outperforms the baseline by a significant margin (10 times) at $1/16$ image size, even achieves the better performance.

Qualitative Result. To further understand the behavior of the learned model, we show the qualitative results in Fig. 4. By observing this figure, we find that our proposed model can ground the phrase in the corresponding subtitle. In the indoor set, there may have multiple references in one sub-

title. It is shown in Fig. 4 that the predicted viewpoint is successfully find the coach, TV and table.

Conclusion

We introduce a new NFoV grounding task which aims to ground subtitles (i.e., natural language phrases) into corresponding views in each $360°$ video. To tackle this task, we propose a novel model with two main innovations: (1) train the network with both of relevant and irrelevant soft-attention visual features, and (2) apply NFoV proposal in feature space for saving time-consuming, memory usage, and making end-to-end training feasible. We achieve the best performance on both recall and precision measurement. In the future, we plan to extend the dataset and model to joint story-telling and NFoV grounding in $360°$ touring videos.

Acknowledgments

References

Assens, M.; McGuinness, K.; Giro, X.; and O'Connor, N. E. 2017. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *ICCV Workshop*.

Chen, J., and Carr, P. 2015. Mimicking human camera operators. In *WACV*.

Chen, J.; Le, H. M.; Carr, P.; Yue, Y.; and Little, J. J. 2016. Learning online smooth predictors for realtime camera planning using recurrent decision trees. In *CVPR*.

Christianson, D. B.; Anderson, S. E.; He, L.-w.; Salesin, D. H.; Weld, D. S.; and Cohen, M. F. 1996. Declarative camera control for automatic cinematography. In *AAAI*.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS Workshop*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Elson, D. K., and Riedl, M. O. 2007. A lightweight intelligent virtual cinematography system for machinima production. In *AIIDE*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*.

Hu, R.; Xu, H.; Rohrbach, M.; Feng, J.; Saenko, K.; and Darrell, T. 2016. Natural language object retrieval. In *CVPR*.

Hu, H.-N.; Lin, Y.-C.; Liu, M.-Y.; Cheng, H.-T.; Chang, Y.-J.; and Sun, M. 2017. Deep 360 pilot: Learning a deep agent for piloting through 360deg sports videos. In *CVPR*.

Johnson, J.; Krishna, R.; Stark, M.; Li, L.-J.; Shamma, D.; Bernstein, M.; and Li, F.-F. 2015. Image retrieval using scene graphs. In *CVPR*.

Judd, T.; Ehinger, K.; Durand, F.; and Torralba, A. 2009. Learning to predict where humans look. In *ICCV*.

As you can see we have the **sectional couch**, the **TV** and **TV stand**.

As you can see we have the **sectional couch**, the **TV** and **TV stand**.

You have your **coffee table** and **end table** and you're also able to paint your **walls** in your common area.

Miyajima island is very famous for its **wooden crafts**.

This is a **rice scope**, **right scope** is famous in japan.

That's why we can see many **rice scope** are being sold.

In the **laundry room** you are actually going to have a full size **washer dryer**.

All of the stacked **washer** and **dryers** that you can see.

Go ahead and not only use it for all of your **laundry** detergent.

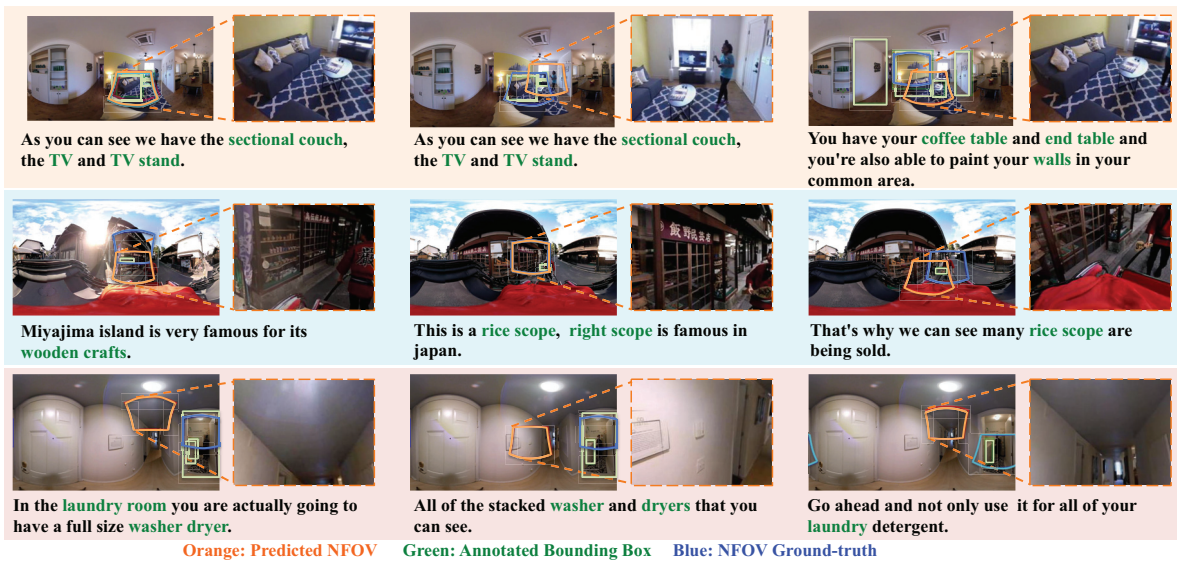**Orange: Predicted NFOV   Green: Annotated Bounding Box   Blue: NFOV Ground-truth**

Figure 4: Qualitative Results on our narrated 360° videos dataset. Top: indoor results. Middle: outdoor results. Bottom: failure case.

Karpathy, A., and Li, F.-F. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Karpathy, A.; Joulin, A.; and Li, F.-F. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*.

Kingma, D., and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Lai, W.; Huang, Y.; Joshi, N.; Buehler, C.; Yang, M.; and Kang, S. B. 2017. Semantic-driven generation of hyperlapse from 360° video. In *TVCG*.

Li, Z.; Tao, R.; Gavves, E.; Snoek, C. G. M.; and Smeulders, A. W. 2017. Tracking by natural language specification. In *CVPR*.

Li, Y.; Fathi, A.; and Rehg, J. M. 2013. Learning to predict gaze in egocentric video. In *ICCV*.

Lin, D.; Fidler, S.; Kong, C.; and Urtasun, R. 2014a. Visual semantic search: Retrieving videos via complex textual queries. In *CVPR*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014b. Microsoft coco: Common objects in context. In *ECCV*.

Lin, Y.-C.; Chang, Y.-J.; Hu, H.-N.; Cheng, H.-T.; Huang, C.-W.; and Sun, M. 2017. Tell me where to look: Investigating ways for assisting focus in 360° video. In *CHI*.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *CVPR*.

Mindek, P.; Čmolík, L.; Viola, I.; Gröller, E.; and Bruckner, S. 2015. Automatized summarization of multiplayer games. In *CCG*.

Paszke, A., and Chintala, S. Pytorch: https://github.com/apaszke/pytorch-dist.

Patney, A.; Kim, J.; Salvi, M.; Kaplanyan, A.; Wyman, C.; Benty, N.; Lefohn, A.; and Luebke, D. 2016. Perceptually-based foveated virtual reality. In *SIGGRAPH*.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Rohrbach, A.; Rohrbach, M.; Hu, R.; Darrell, T.; and Schiele, B. 2016. Grounding of textual phrases in images by reconstruction. In *ECCV*.

Rong, X.; Yi, C.; and Tian, Y. 2017. Unambiguous text localization and retrieval for cluttered scenes. In *CVPR*.

Su, Y.-C., and Grauman, K. 2017. Making 360 video watchable in 2d: Learning videography for click free viewing. In *CVPR*.

Su, Y.-C.; Jayaraman, D.; and Grauman, K. 2016. Pano2vid: Automatic cinematography for watching 360° videos. In *ACCV*.

Sun, X.; Foote, J.; Kimber, D.; and Manjunath, B. 2005. Region of interest extraction and virtual camera control based on panoramic video capturing. *ToM*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Wang, L.; Li, Y.; and Lazebnik, S. 2016. Learning deep structure-preserving image-text embeddings. In *CVPR*.

Xiao, F.; Sigal, L.; and Jae Lee, Y. 2017. Weakly-supervised visual grounding of phrases with linguistic structures. In *CVPR*.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Yu, H., and Siskind, J. M. 2013. Grounded language learning from video described with sentences. In *ACL*.

Zeng, K.-H.; Chen, T.-H.; Niebles, J. C.; and Sun, M. 2016. Title generation for user generated videos. In *ECCV*.

Zeng, K.-H.; Chen, T.-H.; Chuang, C.-Y.; Liao, Y.-H.; Niebles, J. C.; and Sun, M. 2017. Leveraging video descriptions to learn video question answering. In *AAAI*.

Zhang, Y.; Yuan, L.; Guo, Y.; He, Z.; Huang, I.-A.; and Lee, H. 2017. Discriminative bimodal networks for visual localization and detection with natural language queries. In *CVPR*.