Integrating Human Gaze into Attention for Egocentric Activity Recognition

Kyle Min Jason J. Corso
University of Michigan
Ann Arbor, MI 48109
{kylemin,jjcorso}@umich.edu

Abstract

It is well known that human gaze carries significant information about visual attention. However, there are three main difficulties in incorporating the gaze data in an attention mechanism of deep neural networks: (i) the gaze fixation points are likely to have measurement errors due to blinking and rapid eye movements; (ii) it is unclear when and how much the gaze data is correlated with visual attention; and (iii) gaze data is not always available in many real-world situations. In this work, we introduce an effective probabilistic approach to integrate human gaze into spatiotemporal attention for egocentric activity recognition. Specifically, we represent the locations of gaze fixation points as structured discrete latent variables to model their uncertainties. In addition, we model the distribution of gaze fixations using a variational method. The gaze distribution is learned during the training process so that the ground-truth annotations of gaze locations are no longer needed in testing situations since they are predicted from the learned gaze distribution. The predicted gaze locations are used to provide informative attentional cues to improve the recognition performance. Our method outperforms all the previous state-of-the-art approaches on EGTEA, which is a large-scale dataset for egocentric activity recognition provided with gaze measurements. We also perform an ablation study and qualitative analysis to demonstrate that our attention mechanism is effective.

1. Introduction

It has recently been shown that attention mechanisms can boost the performance of neural networks in various tasks by learning to focus on relatively important and salient parts of input signals. Most notably, attention-based recurrent neural networks have achieved great success in machine translation [1, 23] and image captioning [38]. Attention mechanisms have also been widely adopted by deep convolutional neural networks (CNNs) in several forms of feature re-weighting such as spatial attention [37, 27], channel attention [11, 40], etc [35, 33]. These methods usually let neural networks learn *what and where* to focus on from their own responses.

In this paper, we introduce an effective probabilistic method for integrating human gaze into a spatiotemporal attention mechanism. It has been well discussed in cognitive science that human gaze is closely related to a person's behavioral intention and visual attention [34, 4, 9, 28]. At the same time, however, there is always uncertainty in the process of recording the gaze fixation points because of saccadic suppression[1][16] and measurement errors. Furthermore, it is not always guaranteed that the surrounding region around the point of gaze fixation has the most important information, especially when interacting with multiple objects or under dissociation[2][2, 14].

To address such problems, we present a probabilistic modeling method as follows: First, we propose to represent the locations of gaze fixation points in space and time as structured discrete latent variables to model their uncertainties. Second, we model the distribution of the gaze fixations using a variational method. During the training process, the distribution of gaze fixations is learned using the ground-truth annotations of gaze points. Specifically, we propose to reformulate the discrete training objective so that it can be optimized using an unbiased gradient estimator. The gaze locations are predicted from the learned gaze distribution so that the ground-truth annotations of gaze fixation points are no longer needed in testing scenarios. The predicted gaze locations are integrated into a soft attention mechanism to make the intermediate features more attended to informative regions. It is empirically shown that our gaze-combined attention mechanism leads to a significant improvement of activity recognition performance on egocentric videos by providing additional cues across space and time.

We demonstrate the effectiveness of our method on EGTEA [17] and GTEA gaze+ [18], which are large-scale

[1] phenomenon in which visual information is not processed while blinking or under rapid eye movements.

[2] dissociation of the focus of attention is a phenomenon where the points of gaze fixation are not correlated with the visual attention within the field of view.

datasets for egocentric activities provided with gaze measurements. Our method significantly outperforms all the previous state-of-the-art approaches. We also perform an ablation study to verify that probabilistic modeling of gaze data is truly beneficial. We then visualize the spatiotemporal responses of our networks to qualitatively show that the gaze-combined soft attention provides informative attentional cues.

2. Related work

Recently, attention-based recurrent neural networks have been widely adopted for neural machine translation [1, 23] as well as for image captioning [38]. They generate attention vectors by manipulating hidden states of recurrent neural networks and annotated information. Attention mechanisms have also been incorporated with deep CNNs to improve the representation quality of intermediate features by refining the features [37, 27, 11, 40]. They usually introduce attention modules which find channel-wise or spatial-wise attention maps from the average-pooled features descriptors. There are more recent works which utilize both attention methods across spatial and channel dimensions [35, 33]. These methods also have shown that using both average-pooling and max-pooling in parallel is beneficial to building attention maps.

There have been a few attempts to utilize human gaze data for egocentric activity recognition [8, 12, 17]. Fathi et al. [8] propose a conditional generative model that jointly predicts gaze locations and egocentric activity labels. More related and recent works [12, 17] have shown that incorporating gaze data into an attention mechanism can boost the performance of CNNs on egocentric activity recognition. Huang et al. [12] propose Mutual Context Network (MCN) that tries to use human gaze for recognizing activities and use the activity labels for predicting gaze locations. However, MCN has multiple sub-modules that should be trained separately. Furthermore, an inference procedure requires many iterations because of the complicated network architecture. They also use saccades as ground-truth gaze points, which should be ignored to improve the prediction performance. Li et al. [17] is built on a similar probabilistic framework to ours; however, there are three crucial differences. First, to model the distribution of gaze points for T time steps, they use T independent 2D latent variables. This totally ignores the temporal correlation of the gaze distribution, which limits the recognition performance. Second, they use the approximated Gumbel-Softmax objective [13, 25] that introduces a significant bias to a gradient estimator. As a result, the recognition performance of their method is further limited. Third, they directly apply the sampled gaze points $\mathbf{z}^*$ to the input feature map without any modifications. This is vulnerable to situations where the gaze points are misleading and not informative. On the contrary, we use structured discrete latent variables to model the gaze distribution in a 3D space. We apply the direct optimization method to handle this structured latent space, which also minimizes the bias. Moreover, we use the sigmoid activated linear mapping on the sampled gaze points to produce a soft attention map.

3. Background: Direct optimization

Direct optimization [21] was originally proposed for learning a variational auto-encoder (VAE) with discrete latent variables. The objective of VAE is given by:

$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{\mathbf{z}\sim q_\phi}[\log p_\theta(\mathbf{x}|\mathbf{z})] + D_{\text{KL}}[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})] \quad (1)$$

where $\mathbf{x}$ is an input and $\mathbf{z}$ is a discrete latent variable. Computing the expected log-likelihood requires drawing samples from the discrete distribution $q_\phi(\mathbf{z}|\mathbf{x})$, which makes it difficult to optimize. Gumbel-Softmax reparameterization technique [13, 25] was recently suggested to relax the discrete variables to continuous counterparts. However, this continuous relaxation is known to introduce a significant bias when evaluating gradients and become intractable under the high-dimensional structured latent spaces. The direct optimization method introduces an unbiased gradient estimator for the discrete VAE that can be used even under the high-dimensional structured latent spaces. For simplicity, let us rewrite the log-probabilities as follows: $h_\phi(\mathbf{x}, \mathbf{z}) = \log q_\phi(\mathbf{z}|\mathbf{x})$, $f_\theta(\mathbf{x}, \mathbf{z}) = \log p_\theta(\mathbf{x}|\mathbf{z})$. By using the Gumbel-Max trick [26], the expected log-likelihood can be reformulated as follows: $\mathbb{E}_{\mathbf{z}\sim q_\phi}[\log p_\theta(\mathbf{x}|\mathbf{z})] = \mathbb{E}_{\gamma\sim G}[f_\theta(\mathbf{x}, \mathbf{z}^*)]$ where $\mathbf{z}^* = \text{argmax}_{\hat{\mathbf{z}}}\{h_\phi(\mathbf{x}, \hat{\mathbf{z}}) + \gamma(\hat{\mathbf{z}})\}$, G denotes a Gumbel distribution, and $\gamma(\hat{\mathbf{z}})$ represents a random variable sampled from the Gumbel distribution that is associated with each input $\hat{\mathbf{z}}$. Then, the proposed gradient estimator for the expectation term is given in the following form:

$$\nabla_\phi \mathbb{E}_{\gamma\sim G}[f_\theta(\mathbf{x}, \mathbf{z}^*)] = \lim_{\epsilon\to 0} \frac{1}{\epsilon}\Big(\mathbb{E}_{\gamma\sim G}\big[\nabla_\phi h_\phi(\mathbf{x}, \mathbf{z}^*(\epsilon)) - \nabla_\phi h_\phi(\mathbf{x}, \mathbf{z}^*)\big]\Big) \quad (2)$$

where $\mathbf{z}^*(\epsilon) = \text{argmax}_{\hat{\mathbf{z}}}\{\epsilon f_\theta(\mathbf{x}, \hat{\mathbf{z}}) + h_\phi(\mathbf{x}, \hat{\mathbf{z}}) + \gamma(\hat{\mathbf{z}})\}$. The suggested gradient estimator is unbiased when the perturbation parameter ϵ goes to 0, but small ϵ brings a large variance of the estimation. Therefore, in practice, we set ϵ to a large value in the beginning of the training process and decrease it progressively.

4. Method

We start this section by building a probabilistic framework and the loss function of our method. Next, we propose a 3D gaze modeling approach using structured discrete latent variables. We then introduce the direct loss minimiza-

tion approach [21] that is used for optimization in the presence of the structured discrete latent variables. Finally, we describe our overall network architecture for activity recognition that integrates the gaze information into attention.

4.1. Probabilistic framework

Let us consider a recognition task of predicting activity labels $\mathbf{y}$ given an input clip of egocentric videos $\mathbf{x}$, which is equivalent to finding a conditional probability $p(\mathbf{y}|\mathbf{x})$. We represent the gaze locations in space and time with a discrete latent variable $\mathbf{z}$. Then, the conditional probability is written as follows by the law of total probability:

$$p_\theta(\mathbf{y}|\mathbf{x}) = \int p_\theta(\mathbf{y}|\mathbf{x},\mathbf{z})p_\theta(\mathbf{z}|\mathbf{x})d\mathbf{z} \tag{3}$$

where θ denotes the parameters of a network for recognition. Since $\mathbf{z}$ generally has an intractable posterior distribution, we upper bound the negative log-likelihood by taking the negative log on both sides of Equation (3) and introducing the variational approximation $q_\phi(\mathbf{z}|\mathbf{x})$ for gaze modeling as follows:

$$\begin{aligned} -\log p_\theta(\mathbf{y}|\mathbf{x}) &\leq \int -q_\phi \log\Big(p_\theta(\mathbf{y}|\mathbf{x},\mathbf{z})\frac{p_\theta(\mathbf{z}|\mathbf{x})}{q_\phi}\Big)d\mathbf{z} \\ &= -\mathbb{E}_{\mathbf{z}\sim q_\phi}[\log p_\theta(\mathbf{y}|\mathbf{x},\mathbf{z})] + D_{\mathrm{KL}}[q_\phi||p_\theta(\mathbf{z}|\mathbf{x})] \end{aligned} \tag{4}$$

where ϕ denotes parameters of a network for gaze modeling. We use the upper bound in Equation (4) as our loss function.

4.2. Reformulating the training objective

In order to compute the expected log-likelihood of the loss function in Equation (4), we need to sample the gaze points from q_ϕ. We apply the Gumbel-Max trick [26] that is an efficient method of drawing samples from a discrete distribution. For simplicity, let us rewrite the log-probability as follows: $h_\phi(\mathbf{x},\mathbf{z}) = \log q_\phi(\mathbf{z}|\mathbf{x})$. Then, we can draw a gaze sample z^* using the following equation:

$$\mathbf{z}^* = \underset{\hat{\mathbf{z}}}{\mathrm{argmax}}\{h_\phi(\mathbf{x},\hat{\mathbf{z}}) + \gamma(\hat{\mathbf{z}})\} \tag{5}$$

where $\gamma(\hat{\mathbf{z}})$ represents a random variable sampled from a Gumbel distribution that is associated with each input $\hat{\mathbf{z}}$. However, $\mathbf{z}^*$ includes a non-differentiable operation, argmax, so we cannot evaluate the gradient of the expectation term with respect to ϕ using a standard backpropagation algorithm. Here, we propose to apply the direct optimization method [21] to optimize the expected log-likelihood term. In the following, we demonstrate that our loss function can be optimized using the direct optimization method.

Since our task is to classify activity labels, we can model $\mathbf{y}$ given $\mathbf{x}$ and $\mathbf{z}$ with a categorical distribution. Specifically, let us say that there are C number of predefined activity classes. Then, $p_\theta(\mathbf{y}|\mathbf{x},\mathbf{z}) = \prod_{c=1}^{C} p_c^{\mathbb{1}_{y=c}}$ for some class-wise probabilities p_c's that are dependent on $\mathbf{x}$ and $\mathbf{z}$ where $\mathbb{1}_{y=c}$ is an indicator function that is equal to 1 if $y=c$ and 0 otherwise. This allows us to rewrite $\log p_\theta(\mathbf{y}|\mathbf{x},\mathbf{z})$ in the following form:

$$\log p_\theta(\mathbf{y}|\mathbf{x},\mathbf{z}) = \sum_{c=1}^{C} \mathbb{1}_{y=c}\, f_\theta^c(\mathbf{x},\mathbf{z}) \tag{6}$$

where $f_\theta^c(\mathbf{x},\mathbf{z})$'s are the corresponding class-wise log-probabilities. Now, we propose to reformulate the expected log-likelihood using the class-wise log-probabilities:

$$\begin{aligned} \mathbb{E}_{\mathbf{z}\sim q_\phi}[\log p_\theta] &= \sum_{\mathbf{z}} \Big(\mathbb{P}_{\gamma\sim G}[\mathbf{z}^* = \mathbf{z}] \sum_{c=1}^{C} \mathbb{1}_{y=c}\, f_\theta^c(\mathbf{x},\mathbf{z})\Big) \\ &= \sum_{c=1}^{C} \mathbb{1}_{y=c}\, \mathbb{E}_{\gamma\sim G}[f_\theta^c(\mathbf{x},\mathbf{z}^*)] \end{aligned} \tag{7}$$

where G denotes the Gumbel distribution. In Equation (7), We show that the expected log-likelihood can be decomposed into a sum of multiple expectation terms of the class-wise log-probabilities, each multiplied by an indicator function. Since the gradient is a linear operator, we can estimate the gradient of the expected log-likelihood as follows:

$$\nabla_\phi\, \mathbb{E}_{\mathbf{z}\sim q_\phi}[\log p_\theta] = \sum_{c=1}^{C} \mathbb{1}_{y=c}\, \nabla_\phi\, \mathbb{E}_{\gamma\sim G}[f_\theta^c(\mathbf{x},\mathbf{z}^*)] \tag{8}$$

where each class-wise gradient estimator $\nabla_\phi\, \mathbb{E}_{\gamma\sim G}[f_\theta^c(\mathbf{x},\mathbf{z}^*)]$ is computed by applying the direct optimization:

$$\begin{aligned} \nabla_\phi\, \mathbb{E}_{\gamma\sim G}[f_\theta^c(\mathbf{x},\mathbf{z}^*)] = \lim_{\epsilon\to 0} \frac{1}{\epsilon}\Big(&\mathbb{E}_{\gamma\sim G}\big[\nabla_\phi\, h_\phi(\mathbf{x},\mathbf{z}^*(\epsilon,c)) \\ &- \nabla_\phi\, h_\phi(\mathbf{x},\mathbf{z}^*)\big]\Big) \end{aligned} \tag{9}$$

when $\mathbf{z}^*(\epsilon,c) = \mathrm{argmax}_{\hat{\mathbf{z}}}\{\epsilon f_\theta^c(\mathbf{x},\hat{\mathbf{z}}) + h_\phi(\mathbf{x},\hat{\mathbf{z}}) + \gamma(\hat{\mathbf{z}})\}$. Other gradients, such as the gradient of the expected log-likelihood with respect to θ, are obtained using a standard backpropagation algorithm. As a result of the reformulation, we can optimize the training objective without introducing a bias of gradient estimator.

4.3. Structured gaze modeling

We propose to use structured discrete latent variables to model the gaze locations as follows. First, we will write $\mathcal{Z}$ to denote a set of every possible $\mathbf{z}$. Let us say that we want to model the gaze locations in a 3D space: $\mathcal{Z} = \mathbb{R}^{T\times H\times W}$ where T is the length of the temporal dimension and H and W represent the height and width of spatial dimensions. For each time step, gaze is fixated at a single location of a

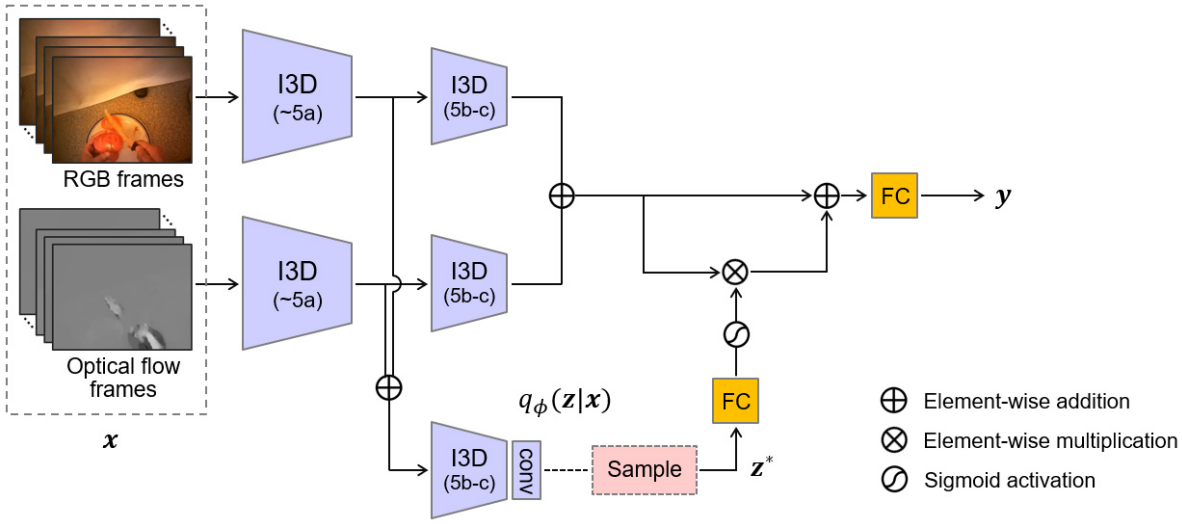


Figure 1: An illustration of our overall network architecture. We use the two-stream I3D [3] as a backbone network. To model the gaze distribution $q_\phi(\mathbf{z}|\mathbf{x})$, we use the same convolutional blocks of the I3D (`Mixed_5b-c`) and add three convolutional layers (conv) on top of it. The two intermediate features at the end of the 4th max-pooling layer (`MaxPool_5a`) are added in an element-wise fashion and used as input to the network for gaze modeling. The sampled gaze point is applied with a fully-connected layer (FC) and with the sigmoid function to produce a soft attention map.

$H \times W$ dimensional space. Therefore, it is more reasonable to represent the gaze locations with a sequence of 2D discrete random variables rather than with a single 3D random variable. Specifically, we assign a 2D discrete random variable to each time step: $\mathbf{z} = (\mathbf{z}_1, ..., \mathbf{z}_t, ..., \mathbf{z}_T)$ where each $\mathbf{z}_t$ is one-hot encoded. For example, if the gaze is fixated at (h, w) on the t-th time step, $\mathbf{z}_t(j, k) = 1$ if $(j, k) = (h, w)$ and 0 otherwise.

Computing $\mathbf{z}^*(\epsilon, c)$ in Equation (9) requires evaluating $f^c_\theta(\mathbf{x}, \mathbf{z})$ for every $\mathbf{z}$, which causes serious overhead. Although our structured gaze modeling reduces the number of possible realizations from 2^{THW} to $(HW)^T$, it is still computationally expensive. We propose to further reduce the number of computations by applying a low-dimensional approximation as suggested by Lorberbom et al. [21]. In particular, we approximate $f^c_\theta(\mathbf{x}, \mathbf{z}) = \sum_{t=1}^{T} f^c_t(\mathbf{x}, \mathbf{z}_t; \theta)$ where $f^c_t(\mathbf{x}, \mathbf{z}_t; \theta) = f^c_\theta(\mathbf{x}, \mathbf{z}^*_1, ..., \mathbf{z}_t, ...\mathbf{z}^*_T)$. This low-dimensional approximation further reduces the number of possible realizations from $(HW)^T$ to THW. We implement the realization of $\mathbf{z}$ by using the batch operation so that we can obtain $\mathbf{z}^*(\epsilon, c)$ in a single forward pass.

4.4. Network architecture

The overall network architecture is illustrated in Figure 1. As a backbone network, we use the two-stream I3D [3] which is a popular network for activity recognition tasks (#Params: 24.7M, FLOPs: 80.2G). To model the gaze distribution $q_\phi(\mathbf{z}|\mathbf{x})$, we use the same convolutional blocks of the I3D (`Mixed_5b-c`) and add three convolutional layers (kernel size=[(1,3,3), (1,3,3), (1,1,1)], stride=[(1,1,1), (1,1,1), (1,1,1)]) on top of it. We add the two intermediate features at the end of the 4th max-pooling layer (`MaxPool_5a`) and use the added feature map as an input to the network for gaze modeling. We draw a sample $\mathbf{z}^*$ using the Equation (5), which is then applied with a fully connect layer and the sigmoid function to produce a soft attention map. The two features at the end of the 5th convolutional block (`Mixed_5c`) are added in an element-wise way, and we apply the soft attention map to the added feature map via a residual connection. Our final network has #Params: 31.9M, FLOPs: 81.3G.

5. Experiments

We evaluate our method on EGTEA [17], which is a large-scale dataset with over 10k video clips of 106 fine-grained egocentric activities and annotated gaze fixations. It is demonstrated that our method outperforms other previous state-of-the-art approaches. Furthermore, we provide a qualitative analysis by visualizing the spatiotemporal responses of our network. We perform additional experiments on GTEA Gaze+ [18] that consists of 2k videos with 44 activity categories.

5.1. Implementation details

Training/testing process. First, we resize each frame to 256×340 and generate optical flow frames by using the TV-L1 algorithm [39]. Following the previous works on the EGTEA dataset [12, 17], we use the I3D pre-trained

Method	Backbone network	Acc (%)	Acc* (%)
Li et al. [17]	I3D [3]	53.30	-
Sudhakaran et al. [32]	ResNet34+LSTM [10, 36]	-	60.76
LSTA [31]	ResNet34+LSTM [10, 36]	-	61.86
MCN [12]	I3D [3]	55.63	-
Kapidis et al. [15]	MFNet [7]	59.44	66.59
Lu et al. [22]	I3D [3]	60.54	68.60
Ours	I3D [3]	**62.84**	**69.58**

Table 1: Performance comparison of our method with other state-of-the-art methods on EGTEA dataset [17]. We report both Acc (mean class accuracy) and Acc* (ratio of correctly classified videos to the total number of videos). Acc is typically lower than Acc* due to an imbalanced class distribution of the dataset.

Method	Backbone network	Acc (%)	Acc* (%)
Sudhakaran et al. [32]	ResNet34+LSTM [10, 36]	-	60.13
MCN [12]	I3D [3]	61.14	-
Ma *et al.* [24]	FCN32s+CNN-M-2048 [20, 5]	-	66.40
Shen *et al.* [30]	SSD+LSTM [19]	-	67.10
Ours	I3D [3]	**64.81**	**68.67**

Table 2: Performance comparison on the GTEA Gaze+ [18] dataset. We report both Acc (mean class accuracy) and Acc* (ratio of correctly classified videos to the total number of videos). Ours again achieves the best performance.

on Kinetics dataset [3] as a backbone network. During the training process, we randomly sample 24-frame input segments and randomly crop 224×224 regions for each segment. We train our network in an end-to-end manner with a batch size of 24 on 8299 training video clips using the first split of the dataset. We use the SGD algorithm with 0.9 momentum and 0.00004 weight decay. The learning rate starts at 0.032 and decays two times by a factor of 10 after 8k and 15k iterations. ϵ is set to 1000 in the beginning and decreases exponentially with a 0.001 annealing rate. We set the minimal ϵ to be 0.1. ϵ goes to this minimum value within 10k iterations. The whole training process of 18K iterations takes less than 12 hours using 4 GPUs (TITAN Xp). For the evaluation, we divide each testing video into non-overlapping 24-frame segments. The whole evaluation process takes less than a half hour using a single GPU.

Dimensions of the latent space. For better comparison, we decided to follow the previous approaches for the dimensions of the latent space. Li et al. [17] suggests predicting gaze points for every 8 frames using the fact that a common duration of gaze fixation is roughly the same as the time interval of 8 frames (about 300ms). It is also suggested to reduce the spatial dimensions of the space for gaze distribution by a factor of 32. This is reasonable since our final goal is to improve the recognition performance, not to predict the exact gaze location in a high-dimensional space. As a result, the dimensions of the 3D latent space for gaze points described in Section 4.3 become $\mathcal{Z} = \mathbb{R}^{3\times7\times7}$ as $T = 24/8$ and $H = W = 224/32$.

5.2. Comparison with the State-of-the-art

We compare our method with other state-of-the-art methods. Performance comparison on the EGTEA dataset is reported in Table 1. We want to point out that Li et al. [17], MCN [12], and Lu et al. [22] use the same backbone network as ours, which is two-stream I3D [3]. Our method outperforms all other methods by a large margin.

We also evaluate our method on the GTEA Gaze+ [18], which is another commonly-used dataset for egocentric activity recognition provided with gaze measurements. It is collected by 6 different human subjects. Following previous works, we perform a leave-one-subject-out cross validation. The performance comparison is reported in Table 2. Our method again achieves the best performance among the recent approaches.

5.3. Qualitative analysis

We visualize the response of the last convolutional layer of our model and of I3D [3] to see how the gaze integration affects the top-down attention of the two networks. We use Grad-CAM++ [6], which is a recently proposed visualization method for CNNs. It is an improved and generalized version of famous Grad-CAM [29]. It is recently shown that

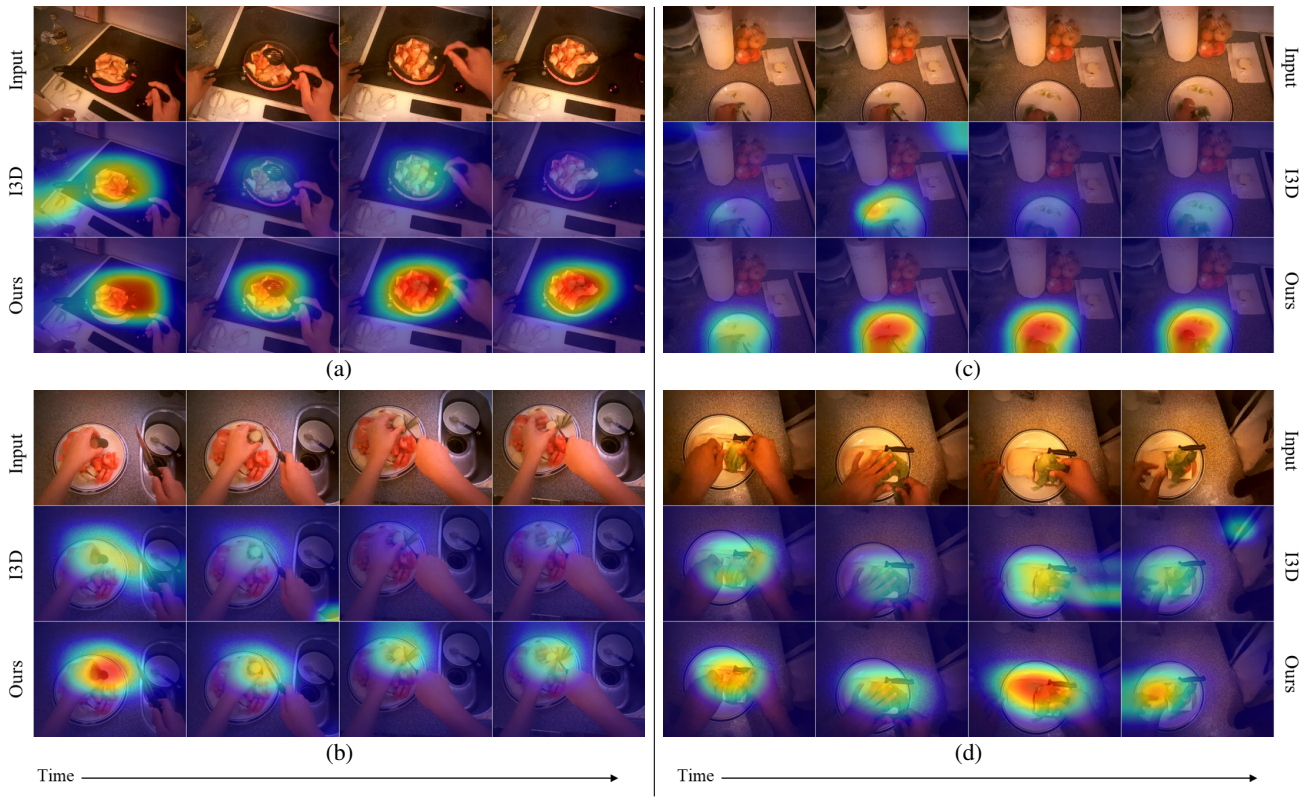


Figure 2: Qualitative results of our model and the baseline network (I3D). We use Grad-CAM++ [6] to visualize the spatiotemporal responses of the last layer of each models. We can observe that our method makes the network better at attending objects or regions which are related to the activity. Activity label of (a): "Move Around bacon", (b): "Cut cucumber", (c): "Cut bell_pepper", (d): "Put lettuce".

Method	Using gaze data during Training	Using gaze data during Testing	Acc (%)	Acc* (%)
I3D w/ Gaze	✓	✓	59.56	67.46
I3D w/ Gumbel-Softmax [13, 25]	✓		61.24	68.69
Ours	✓		**62.84**	**69.58**

Table 3: Performance comparison of different ablative settings. Interestingly, I3D w/ Gaze that uses gaze data also in the testing process performs the worst. The results demonstrate that our structured gaze modeling with direct optimization is effective in improving the performance of egocentric activity recognition. Qualitative analysis regarding this ablation study is provided in the next section.

Grad-CAM++ is effective in understanding 3D CNNs on the task of activity recognition by visualizing the attended locations by the networks across space and time. The visualization results are illustrated in Figure 2. We can clearly observe that our model is better at attending activity-related objects or regions. Specifically, our model is more sensitive to the target objects. The baseline network is sometimes distracted by the background objects. The results qualitatively demonstrate that modeling gaze distributions improves the attentional ability of the networks and the performance of egocentric activity recognition.

5.4. Ablation study

We perform an ablation study on EGTEA dataset [17] as reported in Table 3. "I3D w/ Gaze" refers to the method of using the ground-truth gaze annotations without any gaze modeling. For each input segment, the 3D tensor representing the ground-truth gaze locations $\mathbf{z}_{\text{GT}}$ is first downsampled to have $3 \times 7 \times 7$ dimensions and is applied with a fully-connected layer and the sigmoid function to produce

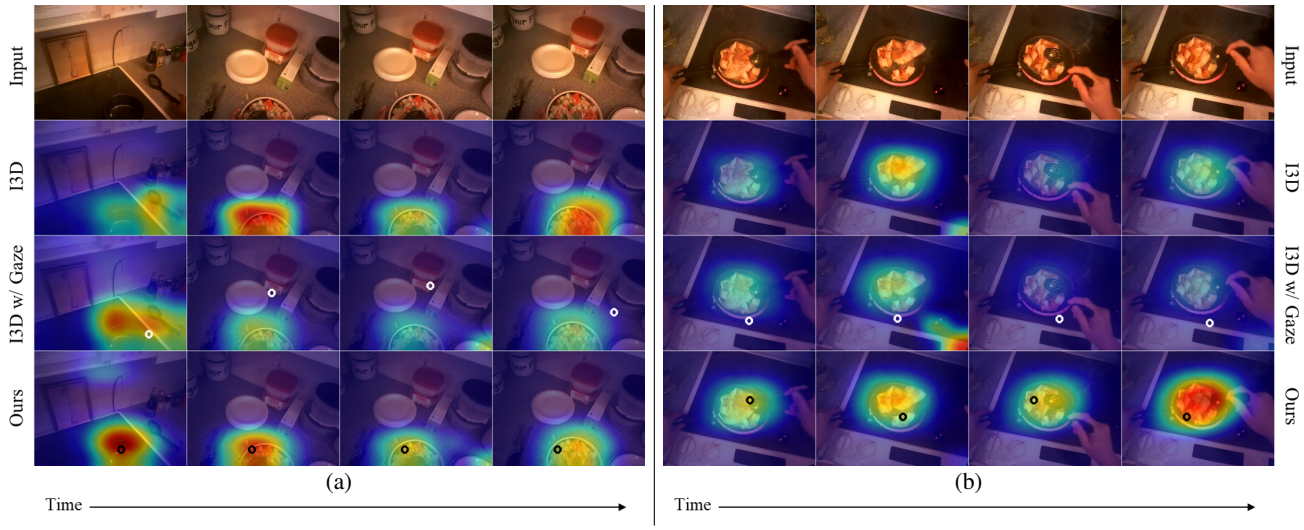


Figure 3: Our method is robust to situations where the ground-truth gaze fixations do not carry activity-related information and are misleading. White marks denote ground-truth annotations of gaze fixations and black marks denote the predicted gaze locations. The predicted gaze locations are successfully fixated on the target objects when the ground-truth annotations are misleading. It demonstrates that our structured gaze modeling with direct optimization is effective. Activity label of (a) is "Mix pasta" and (b) is "Move Around bacon".

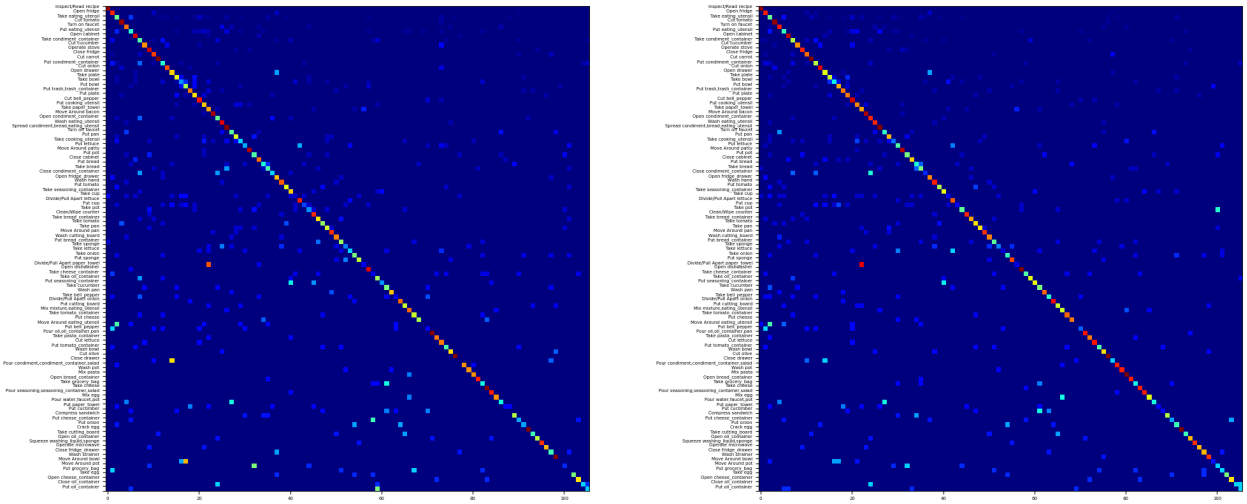
a soft-attention map. This method requires using the gaze data in testing because it does not model the distribution of gaze points. "I3D w/ Gumbel-Softmax [13, 25]" uses the Gumbel-Softmax reparameterization trick to relax the discrete objective to make it continuous. Specifically, it draws a relaxed gaze sample $\mathbf{z}^*_{\text{GS}}$ instead of $\mathbf{z}^*$ in Equation 5 using the following equation: $\mathbf{z}^*_{\text{GS}} = \text{softmax}\big\{\big(h_\phi(\mathbf{x}, \mathbf{z}) + \gamma(\mathbf{z})\big)/\tau\big\}$. We set $\tau = 2$ following the previous work, Li et al. [17], that uses the Gumbel-Softmax objective (but takes different gaze modeling approach). The results indicate that our structured gaze modeling with direct optimization is more effective than the other two methods. Interestingly, "I3D w/ Gaze" that uses gaze data also in the testing process performs the worst. This is probably because some of the ground-truth gaze annotations are not correlated with the actual visual attention. As mentioned in the introduction, measurement error and other uncertainties (saccadic suppression [16] and dissociation [2, 14]) make the annotated gaze points uninformative and sometimes misleading. We argue that our method is capable of learning only the informative gaze distribution that is related to the activities. We qualitatively analyze these interesting results in the following section.

5.5. Robustness to misleading gaze fixations

We perform an additional qualitative analysis to show the robustness of our method to the misleading gaze fixations. Here, misleading gaze points refer to the ground-truth gaze annotations that are not correlated with the actual visual attention. We compare our model with I3D [3] (without any gaze incorporation) and "I3D w/ Gaze" which uses gaze data in training and testing without gaze modeling. We again use Grad-CAM++ [6] to visualize the spatiotemporal activation maps of the last convolutional layer of each model. Figure 3 illustrates the situations where the ground-truth gaze points are not fixated at the activity-related objects or regions. In these examples, the gaze points are not informative and misleading: the ground-truth gaze points are fixated on the background, not on the pan. This leads to blurry and noisy activation maps of "I3D w/ Gaze" because it uses the misleading ground-truth gaze points directly as a soft-attention map. We can observe that our method is robust to such misleading gaze points while "I3D w/ Gaze" is not. Specifically, the predicted gaze locations (denoted as black marks) are successfully fixated on the target objects when the ground-truth annotations (denoted as white marks) are not. It demonstrates the effectiveness of our proposed structured gaze modeling with direct optimization.

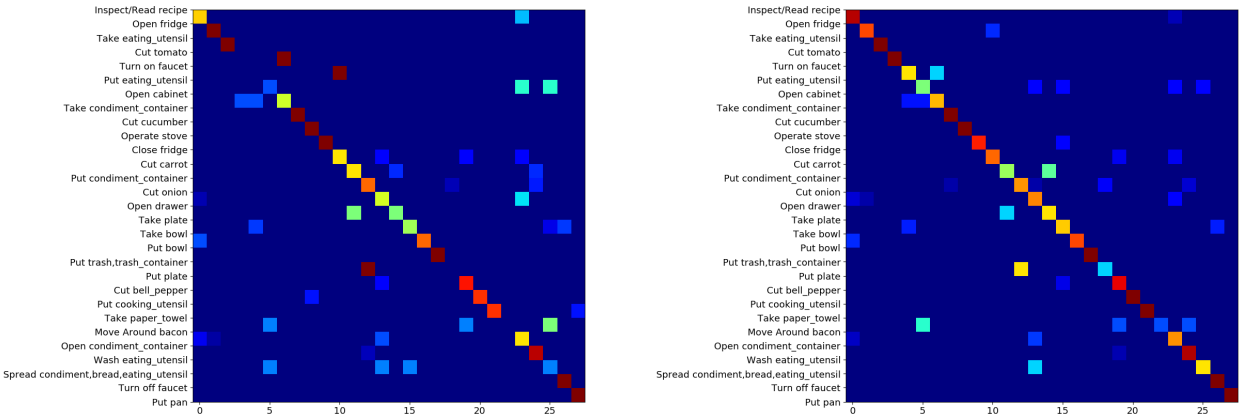
6. Additional analysis

We visualize confusion matrices for the baseline network (I3D [3]) and our method on the EGTEA dataset [17] in Figure 4. Our method outperforms the baseline at least by 0.1% on 28 classes. For better comparison, we also visualize confusion matrices of the two methods on these 28 classes in Figure 5. We can observe that many activities containing "Cut", "Take", and "Put" are benefitted from our gaze incorporation.



(a) Baseline (b) Ours

Figure 4: Confusion matrices for the baseline (I3D [3]) and ours on the EGTEA dataset [17].



(a) Baseline (b) Ours

Figure 5: Confusion matrices for the baseline and ours on 28 classes where our method beats the baseline by a meaningful margin (0.1%). We can observe that many activities containing "Cut", "Take", and "Put" are better recognized by our gaze incorporated model.

7. Conclusion

We have presented an effective method of integrating human gaze into attention on the task of egocentric activity recognition. Incorporating gaze data is non-trivial because there is always uncertainty in the process of recording and the regions near the gaze fixation points are sometimes uninformative. Our method addresses both problems with a probabilistic modeling and an efficient optimization technique. We implement the overall network structures with a simple and powerful 3D CNNs. We evaluate our method in various ways on large-scale datasets. An ablation study demonstrates that incorporating gaze data improves the recognition performance. This is because gaze is correlated with egocentric activity. Moreover, it shows that our proposed structured gaze modeling provides performance improvements by extracting only the informative cues. Interestingly, modeling gaze distribution is more effective in improving the performance than when using ground-truth gaze measurements. We argue that our model is capable of learning only the informative gaze distribution, which is related to the activities of interest. We also qualitatively analyze the effectiveness of our model using the state-of-the-art visualization technique. Our method outperforms all the other previous methods on the task of egocentric activity recognition.

Acknowledgement We thank Ryan Szeto and Christina Jung for their valuable comments. This research was, in part, supported by NIST grant 60NANB17D191.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Julie A Brefczynski and Edgar A DeYoe. A physiological correlate of the'spotlight'of visual attention. *Nature neuroscience*, 2(4):370, 1999.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4724–4733. IEEE, 2017.

[4] Umberto Castiello. Understanding other people's actions: intention and attention. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2):416, 2003.

[5] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[6] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2018.

[7] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Multi-fiber networks for video recognition. In *Proceedings of the european conference on computer vision (ECCV)*, pages 352–367, 2018.

[8] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, pages 314–327. Springer, 2012.

[9] Alexandra Frischen, Andrew P Bayliss, and Steven P Tipper. Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological bulletin*, 133(4):694, 2007.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[12] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and actions. *arXiv preprint arXiv:1901.01874*, 2019.

[13] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[14] Chi-Hung Juan, Stephanie M Shorter-Jacobi, and Jeffrey D Schall. Dissociation of spatial attention and saccade preparation. *Proceedings of the National Academy of Sciences*, 101(43):15541–15544, 2004.

[15] Georgios Kapidis, Ronald Poppe, Elsbeth van Dam, Lucas Noldus, and Remco Veltkamp. Multitask learning to improve egocentric action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[16] Bart Krekelberg. Saccadic suppression. *Current Biology*, 20(5):R228–R229, 2010.

[17] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 619–635, 2018.

[18] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 287–295, 2015.

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[21] Guy Lorberbom, Andreea Gane, Tommi Jaakkola, and Tamir Hazan. Direct optimization through $\arg\max$ for discrete variational auto-encoder, 2018.

[22] Minlong Lu, Danping Liao, and Ze-Nian Li. Learning spatiotemporal attention for egocentric action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[23] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[24] Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1903, 2016.

[25] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[26] Chris J Maddison, Daniel Tarlow, and Tom Minka. A* sampling. In *Advances in Neural Information Processing Systems*, pages 3086–3094, 2014.

[27] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[28] Ann T Phillips, Henry M Wellman, and Elizabeth S Spelke. Infants' ability to connect gaze and emotional expression to intentional action. *Cognition*, 85(1):53–78, 2002.

[29] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[30] Yang Shen, Bingbing Ni, Zefan Li, and Ning Zhuang. Egocentric activity prediction via event modulated attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 197–212, 2018.

[31] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9954–9963, 2019.

[32] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: nailing down object-centric attention for egocentric activity recognition. *arXiv preprint arXiv:1807.11794*, 2018.

[33] Masanori Suganuma, Xing Liu, and Takayuki Okatani. Attention-based adaptive selection of operations for image restoration in the presence of unknown combined distortions. *arXiv preprint arXiv:1812.00733*, 2018.

[34] Joan N Vickers. Advances in coupling perception and action: the quiet eye as a bidirectional link between gaze, attention, and action. *Progress in brain research*, 174:279–288, 2009.

[35] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[36] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

[37] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016.

[38] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[39] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007.

[40] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.