Multimodal Event Causality Reasoning with Scene Graph Enhanced Interaction Network

Jintao Liu, Kaiwen Wei*, Chenglong Liu

University of Chinese Academy of Sciences
{liujintao201, weikaiwen19, liuchenglong20}@mails.ucas.ac.cn

Abstract

Multimodal event causality reasoning aims to recognize the causal relations based on the given events and accompanying image pairs, requiring the model to have a comprehensive grasp of visual and textual information. However, existing studies fail to effectively model the relations of the objects within the image and capture the object interactions across the image pair, resulting in an insufficient understanding of visual information by the model. To address these issues, we propose a **S**cene Graph **E**nhanced **I**nteraction **N**etwork (SEIN) in this paper, which can leverage the interactions of the generated scene graph for multimodal event causality reasoning. Specifically, the proposed method adopts a graph convolutional network to model the objects and their relations derived from the scene graph structure, empowering the model to exploit the rich structural and semantic information in the image adequately. To capture the object interactions between the two images, we design an optimal transport-based alignment strategy to match the objects across the images, which could help the model recognize changes in visual information and facilitate causality reasoning. In addition, we introduce a cross-modal fusion module to combine textual and visual features for causality prediction. Experimental results indicate that the proposed SEIN outperforms state-of-the-art methods on the Vis-Causal dataset.

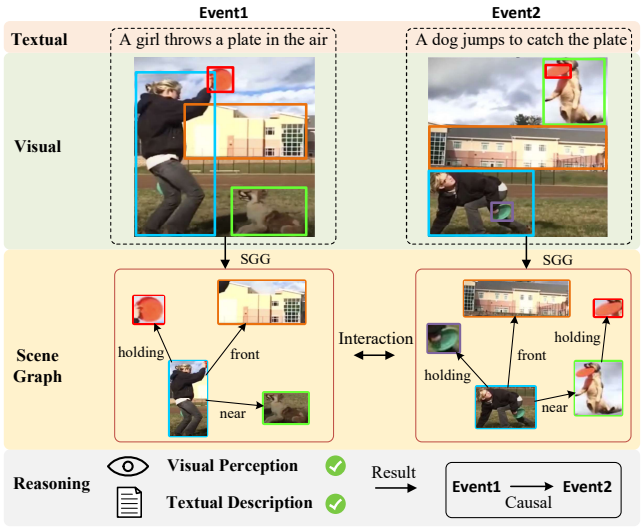


Figure 1: An example of employing scene graph for multimodal event causality reasoning, which could provide rich structural and semantic information for visual understanding of the image.

Introduction

Understanding causality from multimodal daily events is a challenging task and has attracted increasing attention from the community. Take Fig. 1 as an example, the reasoning model should be able to identify the causal relations between the two events based on *A girl throws a plate in the air* and *A dog jumps to catch the plate* as well as their associated image pairs. This task exhibits extensive applications in text and vision domains, including visual commonsense reasoning (Hildebrandt et al. 2020), dense video captioning (Iashin and Rahtu 2020), and machine reading comprehension (Rajani et al. 2019).

Recently, many studies have concentrated on this task. Zhang et al. (2021) have embarked on extracting causal relations from time-consecutive images by incorporating event descriptions and visual context representations. Chadha and Jain (2021) utilized both videos and natural language captions to infer visual-semantic commonsense knowledge with causal rationalization. Afterward, Ma and Tong (2022) combined visual perception and linguistic commonsense to enhance daily events causality reasoning and exploited object features to refine visual perception. Despite promising advancements achieved by existing studies, they still tend to overlook the importance of the following two critical concerns:

(1) **The relations between objects in the image.** Previous works mainly focus on global features or object features of the image, ignoring the significance of modeling the relations between objects. We contend that the relations between paired objects are crucial for understanding the structural and semantic information of the image. As shown in Fig. 1, for the first image, if the model could discern *a girl is holding a plate* and *a dog is near the girl*, it would better comprehend the visual semantic information conveyed by this image. Recently, scene graph generation (SGG), which

*Corresponding author

aims to express objects and relations between objects in the image, has been gradually applied to various vision-based tasks. As a result, adopting SGG to recognize the objects and their relations can foster a more structured understanding of the image. Nevertheless, how to effectively model the objects and their relations remains to be studied.

(2) **The object interactions across the image pair.** Due to the lack of object interactions, the model struggles to recognize the visual information variation between the images. Intuitively, humans can identify the association between two images by observing changes in objects and their relations across the image pair. Inspired by this, it is desirable for our model to capture the interactions of objects from the images. For example in Fig. 1, through object alignment and interaction between the images, the model could understand the changes in visual information from *a girl is holding a plate* in the first image to *a dog is holding a plate* in the second image, thus facilitating the identification of event causalities. However, capturing such interactions is challenging, and directly combining objects in two images might lead to inconsistencies and introduce noise.

To address the above issues, we propose a novel **S**cene Graph **E**nhanced **I**nteraction **N**etwork (SEIN) in this paper, which can leverage the interactions of the generated scene graph for multimodal event causality reasoning. Concretely, we first construct a scene graph for each image to obtain a sufficient structured understanding, where the nodes represent objects or relations, and edges represent the connections between them. Then we employ a Graph Convolutional Network (GCN) to model the objects and their relations to obtain context-aware node embeddings. To capture the object interactions across the two images, we propose optimal transport-based alignment to match the objects in the images, which could recognize the changes in visual information and enhance reasoning capability. And we combine the object features from two images according to the transportation cost matrix. Besides, we adopt a cross-modal fusion module based on the multi-head attention mechanism to integrate textual and visual features for causality prediction. The main contributions of this paper can be summarized as follows:

- This paper proposes a SEIN framework, the first work to our knowledge that exploits the interactions of the scene graph to recognize visual information changes for multimodal event causality reasoning.
- We adopt a GCN architecture to model the objects and their relations in the image, and design an optimal transport-based alignment strategy to capture the object interactions across the image pair.
- Experimental results demonstrate that the proposed SEIN achieves state-of-the-art performance on the Vis-Causal dataset. Further analyses indicate the effectiveness and generalization ability of SEIN.

Related Work

Multimodal Event Causality Reasoning

Previous approaches for event causality identification primarily focus on textual modality (Cao et al. 2021; Wei et al. 2021; Liu et al. 2023b). They seek to leverage external knowledge (Liu, Chen, and Zhao 2020; Cao et al. 2021; Wei et al. 2022) or prompt-tuning technique (Shen et al. 2022; Liu et al. 2023a; Wang et al. 2022b) to identify event causalities. Although these methods have achieved some success, their reliance on a single modality limits their applicability in real-world scenarios. Recently, multimodal event causality reasoning has garnered increasing attention. Zhang et al. (2021) first extract event causalities from time-consecutive images and natural language descriptions using event and visual context representations. Chadha and Jain (2021) utilize both videos and natural language captions to infer visual-semantic commonsense knowledge with causal rationalization. After that, Ma and Tong (2022) leverage visual perception and linguistic commonsense for this task and exploit object features to refine visual perception. However, these methods do not consider the rich structural and semantic information in the scene graph and the interactions of objects between images, making it challenging to adequately exploit the visual information.

Scene Graph Generation

Scene Graph Generation (SGG) has gained substantial interest in the field of computer vision since proposed in Xu et al. (2017). The purpose of SGG is to recognize objects and the relations between paired objects within an image, and then construct a graph where the objects serve as nodes and the relations between them serve as either nodes (Zareian, Karaman, and Chang 2020) or edges (Li, Zhang, and He 2022). It can provide valuable structural and semantic information for a range of downstream tasks, such as image retrieval (Yoon et al. 2021), visual commonsense reasoning (Wang et al. 2022c), and multimodal information extraction (Wang et al. 2022a). Various approaches have emerged to generate scene graphs in different ways (Wang et al. 2019; Lu et al. 2021), and some studies extend SGG from images to videos (Ji et al. 2020; Cong et al. 2021). Nevertheless, since multimodal event causality reasoning focuses on identifying causal relations between two events, it is not feasible to directly introduce SGG for this task.

Optimal Transport

Optimal Transport (OT) mainly studies how to achieve the optimal allocation of resources between two probability distributions, which has a wide range of applications in many areas such as self-supervision learning (Wu et al. 2022), domain adaptation (Xu et al. 2022), and label assignment (Wei et al. 2023). The fundamental idea behind OT is to determine the most efficient way to transform one distribution into another, taking into account the costs associated with the transportation plan. An influential work in OT is the fast solver proposed by Cuturi (2013), which adopted Sinkhorn's matrix scaling algorithm with an entropic regularization term to solve the OT problem orders of magnitude faster than transport solvers. Building upon this, Xie et al. (2019) optimized the Sinkhorn algorithm and proposed an IPOT solver that leveraged an inexact proximal point method with proximal operator approximately evaluated at each iteration. Li et al. (2022) seeks to capture event argument structures with event

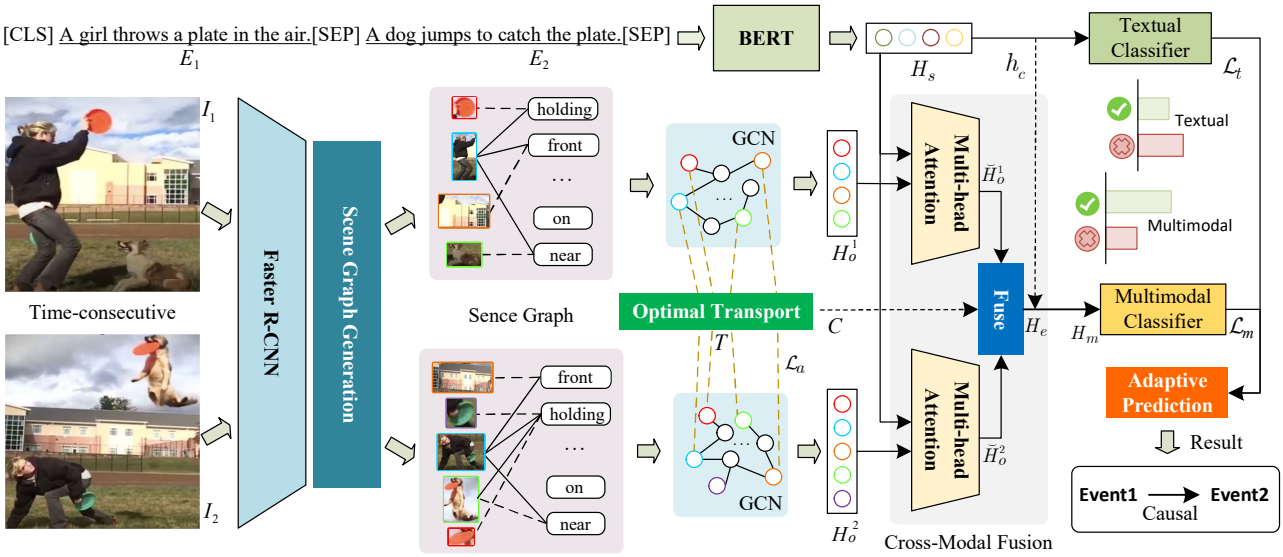


Figure 2: The overview of the proposed SEIN framework.

graph alignment. In this paper, we employ optimal transport to enhance the global alignments and semantic interactions of the scene graphs.

Task Formulation

The goal of multimodal event causality reasoning is to recognize causality between two given events, which contain images cropped from the videos and corresponding natural language descriptions. Following Zhang et al. (2021), this task is formally defined as follows:

(1) The input of the model is two time-consecutive images and candidate event sets. The images $\mathcal{I}$ are cropped from the daily life video at equal time intervals. Each image pair consists of two images $I_1, I_2 \in \mathcal{I}$ in temporal order (i.e., I_1 appears before I_2, and I_1 and I_2 correspond to cause and effect images, respectively). The event set associated with I_1 is denoted as $\mathcal{E}_1$ and the event set encompassing all images sampled from the video is denoted as $\mathcal{E}_v$.

(2) Given an image pair I_1, I_2 and event set $\mathcal{E}_1$, for each $E_1 \in \mathcal{E}_1$, the objective is to find all events $E_2 \in \mathcal{E}_v$ that E_1 causes E_2. The output of the model is a causality score indicating the probability that E_1 leads to the occurrence of E_2 for each $E_2 \in \mathcal{E}_v$.

Methodology

The overview of the proposed method is illustrated in Fig. 2. For the text modality, we concatenate the two event descriptions and encode them into hidden representations with BERT architecture. For the visual modality, we first construct scene graph for each image and employ a GCN architecture to model the objects and relations between paired objects. To capture the object interactions across the two images, we adopt optimal transport-based alignment to match the objects from the scene graphs. Then we combine the text and image representations with a multi-head attention mechanism and integrate the object features based on the cost matrix. Finally, in order to obtain the overall prediction results, we introduce an adaptive prediction strategy to fuse the outputs from textual and multimodal classifiers.

In this section, we first introduce the acquisition of text and visual representations. Then we introduce the optimal transport-based alignment strategy. Subsequently, we present the cross-modal fusion module. Finally, we elucidate model training and prediction processes.

Textual Representation

To acquire a textual comprehension of the events, we first concatenate the two event descriptions with *[SEP]* token and add a *[CLS]* token at the beginning. Then we adopt pre-trained BERT (Devlin et al. 2019) architecture to encode the sequence into hidden states:

$$H_s = \text{BERT}([CLS]\ E_1\ [SEP]\ E_2\ [SEP]) \quad (1)$$

where $H_s \in \mathbb{R}^{n \times d}$, d is the dimension of hidden states. Note that we pre-train BERT with event pairs from ATOMIC (Sap et al. 2019) knowledge base to improve the reasoning ability of the model before fine-tuning.

Scene Graph

In this work, we adopt the technique of SGG to extract objects and relations between the paired objects, which could enable the model to grasp a higher-level visual understanding of the image. Specifically, we first leverage the object detector Faster R-CNN (Ren et al. 2015) pre-trained on Visual Genome (Krishna et al. 2017) to detect a set of objects for each image. Then the public Scene Graph Diagnosis toolkit (Tang et al. 2020) is utilized to recognize relations between each pair of objects.

Formally, for an image, the object set is denoted as $O = \{o_i\}$ and the relation set is denoted as $R = \{r_{ij}\}$, where r_{ij} indicates the relation between objects o_i and o_j. We define

the scene graph as $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N} = O \cup R$ denotes the set of nodes containing objects and relations, $\mathcal{E}$ denotes the set of directed edges. It is worth noting that when r_{ij} exists, we add two directed edges $o_i \rightarrow r_{ij}$ and $r_{ij} \rightarrow o_j$ to $\mathcal{E}$ during the construction process of the edge set.

Visual Representation

For the embedding layer of the graph, we design three types of feature representations for each object and relation: (1) visual feature, (2) position feature, (3) category feature. Specifically, the visual features of objects $f^v_{o_i} \in \mathbb{R}^{d_v}$ are obtained from the region of interest (ROI) features of the object detector and the visual features of relations $f^v_{r_{ij}} \in \mathbb{R}^{d_v}$ are the relation representations before the final prediction layer in the SGG model. The position features of objects $f^p_{o_i} \in \mathbb{R}^{d_p}$ and relations $f^p_{r_{ij}} \in \mathbb{R}^{d_p}$ are converted from bounding box coordinates and union box coordinates, respectively. Besides, the category features of objects $f^c_{o_i} \in \mathbb{R}^{d_c}$ and relations $f^c_{r_{ij}} \in \mathbb{R}^{d_c}$ are obtained from pre-trained Glove word embeddings (Pennington, Socher, and Manning 2014) corresponding to the category labels of objects and relations. Then we fuse the three types of features with a linear layer followed by a ReLU activation function:

$$\begin{aligned} f_{o_i} &= \mathrm{ReLU}(W^v_{o_i} f^v_{o_i} + W^p_{o_i} f^p_{o_i} + W^c_{o_i} f^c_{o_i}) \\ f_{r_{ij}} &= \mathrm{ReLU}(W^v_{r_{ij}} f^v_{r_{ij}} + W^p_{r_{ij}} f^p_{r_{ij}} + W^c_{r_{ij}} f^c_{r_{ij}}) \end{aligned} \tag{2}$$

where $W^v \in \mathbb{R}^{d_v \times d}$, $W^p \in \mathbb{R}^{d_p \times d}$, and $W^c \in \mathbb{R}^{d_c \times d}$ are trainable parameters. The fused object and relation features $f_{o_i} \in \mathbb{R}^d$, $f_{r_{ij}} \in \mathbb{R}^d$ are employed to initialize node embeddings in the graph.

After that, we adopt Graph Convolutional Networks (Kipf and Welling 2017) to aggregate information of neighborhoods and get context-aware representations for objects. Each node in the l-th GCN layer is updated according to the representations of neighbor nodes as:

$$\begin{aligned} F^l_o &= F^{l-1}_o + \mathrm{ReLU}(A_{ro} F^{l-1}_r W^l_r) \\ F^l_r &= F^{l-1}_r + \mathrm{ReLU}(A_{or} F^{l-1}_o W^l_o) \end{aligned} \tag{3}$$

where $F_o = [f_{o_i}] \in \mathbb{R}^{N_o \times d}$ and $F_r = [f_{r_{ij}}] \in \mathbb{R}^{N_r \times d}$, N_o and N_r are the number of objects and relations respectively, $A_{ro} \in \mathbb{R}^{N_o \times N_r}$ and $A_{or} \in \mathbb{R}^{N_r \times N_o}$ are the normalized adjacency matrices from objects to relations, and from relations to objects, $W^l \in \mathbb{R}^{d \times d}$ are trainable parameters of the l-th GCN layer. Finally, we can obtain the object representations $H_o = F^L_o \in \mathbb{R}^{N_o \times d}$ from the output of the L-th GCN layer.

It should be noted that we use the above approach to get visual representations of objects for each image in the pair, which can be denoted as $H^1_o = [h^1_{o_i}] \in \mathbb{R}^{N^1_o \times d}$ and $H^2_o = [h^2_{o_j}] \in \mathbb{R}^{N^2_o \times d}$, respectively.

Optimal Transport-Based Alignment

Since the same object has similar representations in different images, we adopt optimal transport to achieve global alignment and interactions between the object features in the two time-consecutive images, which is beneficial for recognizing visual information changes.

This work seeks to get the minimal OT distance between H^1_o and H^2_o, which is defined as:

$$\mathrm{OTA}(H^1_o,\ H^2_o) = \min_T \langle T, C \rangle \tag{4}$$

where $\langle T, C \rangle = \mathrm{Tr}(T^\top C)$ denotes the Frobenius inner product, $T \in \mathbb{R}^{N^1_o \times N^2_o}$ denotes the transportation plan, C represents the cost matrix between H^1_o and H^2_o. In the implementations, we use the cosine distance between the two objects to compute the cost matrix:

$$C_{ij} = 1 - \frac{h^1_{o_i} \cdot {h^2_{o_j}}^\top}{||h^1_{o_i}||_2 \, ||h^2_{o_j}||_2} \tag{5}$$

To solve Eq 4, we employ the IPOT method (Xie et al. 2019) to calculate the approximated T.

Cross-Modal Fusion

After obtaining textual and visual representations, a cross-modal fusion module is designed to effectively fuse the two modalities. We first adopt a multi-head attention mechanism (Vaswani et al. 2017) to capture interactions between the textual and visual modalities, which can be formulated as:

$$\begin{aligned} \mathrm{Head}_i &= \mathrm{softmax}(\frac{[QW^Q_i][KW^K_i]^\top}{\sqrt{d/h}})[VW^V_i] \\ \mathrm{MHA}(Q,K,V) &= [\mathrm{Head}_1 \oplus \cdots \oplus \mathrm{Head}_h]W_a \end{aligned} \tag{6}$$

where h represents the number of heads, $\oplus$ denotes concatenation operation, $\{W^Q_i, W^K_i, W^V_i\} \in \mathbb{R}^{d \times d/h}$ are trainable parameters. We take the object representations from each image as query, and the textual representations as key and value respectively to obtain representations as:

$$\begin{aligned} \breve{H}^1_o &= \mathrm{MHA}(H^1_o, H_s, H_s) \\ \breve{H}^2_o &= \mathrm{MHA}(H^2_o, H_s, H_s) \end{aligned} \tag{7}$$

Empirically, object pairs exhibiting smaller cosine distances tend to indicate strong object matching or semantic relevance. Therefore, these pairs hold greater significance in mining causal clues. Based on this observation, we select the top-$\mathcal{K}$ object pairs with the lowest cosine distance in cost matrix C and concatenate their features as the fused representations: $H_e = [h^k_e] \in \mathbb{R}^{\mathcal{K} \times 2d}$, where $h^k_e = [\breve{h}^1_{o_i};\ \breve{h}^2_{o_j}] \in \mathbb{R}^{1 \times 2d}$.

Afterward, we seek to aggregate the features with the guide of textual information. We take the representation of *[CLS]* token h_c as the overall textual representation. Then we concatenate h_c and h^i_e and feed them into a fully-connected layer to compute attention score. Finally, we sum the fused representations with the attention score to obtain multimodal representations:

$$\begin{aligned} \alpha_i &= \mathrm{softmax}(W_e[h_c; h^i_e]) \\ H_m &= \sum_{i=1}^{\mathcal{K}} \alpha_i \cdot h^i_e \end{aligned} \tag{8}$$

Algorithm 1: The Training Process of SEIN

Input: Training set $\mathcal{D} = \{(E_1^i, I_1^i), (E_2^i, I_2^i)\}_{i=1}^N$, where E_1 and I_1 represent the text and image of the first event, E_2 and I_2 represent the second.

Training:

for each batch $\mathcal{D}_b \in \mathcal{D}$ **do**
  for any event pair $\in \mathcal{D}_b$ **do**
    Get H_s by Eq. 1;
    Construct scene graph $\mathcal{G}$ for each image;
    Get H_o^1 and H_o^2 by Eq. 2 and Eq. 3;
    OT-based alignment $\mathrm{OTA}(H_o^1,\ H_o^2)$ by Eq. 4;
    Get $\breve{H}_o^1$ and $\breve{H}_o^2$ by Eq. 7;
    Fuse $\breve{H}_o^1$ and $\breve{H}_o^2$ into H_e according to C in Eq. 5;
    Get H_m by Eq. 8;
    Compute classification loss $\mathcal{L}_t$ and $\mathcal{L}_m$;
    Compute object alignment loss $\mathcal{L}_a$;
  end for
  Compute batch loss $\mathcal{L} = \lambda_1\mathcal{L}_t + \lambda_2\mathcal{L}_m + \lambda_3\mathcal{L}_a$;
  Stochastic gradient update model parameters;
end for

Model Training and Prediction

The textual and multimodal representations are fed into fully-connected layers to obtain the predicted causal scores $\hat{y}_t$ and $\hat{y}_m$, respectively. We adopt binary cross-entropy loss as a training objective for textual classification:

$$\mathcal{L}_t = -\frac{1}{N}\sum_{i=1}^{N} y_t^i \log \hat{y}_t^i + (1 - y_t^i)\log(1 - \hat{y}_t^i) \tag{9}$$

Similarly, we use cross-entropy loss to compute multimodal classification loss $\mathcal{L}_m$. We also regard the distance between two scene graphs as a training objective:

$$\mathcal{L}_a = \frac{1}{N}\sum_{i=1}^{N} \mathrm{OTA}(H_o^1,\ H_o^2) \tag{10}$$

The overall training loss can be calculated as:

$$\mathcal{L} = \lambda_1\mathcal{L}_t + \lambda_2\mathcal{L}_m + \lambda_3\mathcal{L}_a \tag{11}$$

The training process of SEIN is summarized in Algorithm 1. In the prediction stage, due to the different contributions of textual and multimodal classification to the outcomes (Ma and Tong 2022), we leverage an adaptive prediction strategy to calculate the causal score. The confidence score is defined as $s = max(\hat{y}, 1-\hat{y})$ to measure the significance of different modalities, and the final predicted causal score is:

$$\hat{y}_f = (1 + \beta^2)\frac{\hat{y}_t \cdot \hat{y}_m}{\beta^2\hat{y}_t + \hat{y}_m} \tag{12}$$

where β is an adaptive weight factor, which is defined as:

$$\beta = \begin{cases} e^{\sqrt{s_m - s_t}}, & s_m > s_t \\ e^{-\sqrt{s_t - s_m}}, & s_m < s_t \end{cases} \tag{13}$$

Split	#Video	#Image Pair	#Event
Train	800	1609	82731
Valid	100	208	10608
Test	100	191	9053

Table 1: Statistics of the Vis-Causal dataset.

Experiments

Experimental Settings

Dataset. We conduct experiments to evaluate our model on the Vis-Causal dataset (Zhang et al. 2021), which is widely used for multimodal daily event causality reasoning. The images in the dataset are collected from YouTube videos, which cover most categories of daily life, i.e., Sports, Socializing, Household, Personal Care, and Eating. Based on the images, the goal is to find the event from the candidate set that has causality with the given event. The statistics of the dataset are listed in Table 1.

Evaluation Metrics. In line with previous works (Zhang et al. 2021; Ma and Tong 2022), we employ Recall@K (R@K) as evaluation metric. R@K reflects the ratio of the correct outcomes in the top-K plausible scores to the total number of ground truth causality events. This paper uses R@1, R@5, and R@10 to evaluate the model performance.

Implementation Details. All experiments are conducted on NVIDIA Tesla V100 GPU with Pytorch framework. We adopt pre-trained BERT-BASE-UNCASED architecture from HuggingFace's Transformers library as textual encoder. We use Faster R-CNN (Ren et al. 2015) pre-trained on Visual Genome to detect objects and leverage the public Scene Graph Diagnosis toolkit (Tang et al. 2020) to identify relations between each pair of objects. The hyper-parameters λ_1, λ_2, and λ_3 are set to 0.5, 0.3, and 0.1, respectively. The number of paired objects $\mathcal{K}$ is set to 10. The number of GCN layers L is set to 2. The model is trained for 25 epochs with a learning rate of 5e-5 and a batch size of 16. The dimension of the hidden representations d is set to 768. We utilize an early stop strategy and Adam optimizer to update model parameters.

Compared Methods

In this work, we compare the proposed SEIN with the following baselines: (1) **Random**, which means randomly selecting an event from the candidate event set as the prediction result. (2) **BERT** (Devlin et al. 2019), a method that leverages the BERT model to encode the textual modality for causality reasoning without considering the visual modality. (3) **VCC** (Zhang et al. 2021), which uses event descriptions and visual context representations to extract causal clues from time-consecutive images. (4) **iReason** (Chadha and Jain 2021), which seeks to infer visual-semantic commonsense knowledge using both videos and natural language captions with causal rationalization. (5) **OARNet** (Ma and Tong 2022), which combines visual perception and linguistic commonsense for daily events causality reasoning and exploits object representations to refine visual perception.

Method	Metrics	Sports	Socializing	Household	Care	Eating	Overall
Random	R@1	0.67	3.64	1.69	0.00	9.09	2.13
	R@5	14.19	16.36	15.25	11.11	27.27	15.25
	R@10	28.38	38.18	27.12	33.33	27.27	30.14
BERT	R@1	12.16	7.27	3.39	0.00	18.18	9.22
	R@5	29.05	32.73	37.29	55.56	54.55	33.33
	R@10	62.84	67.27	49.15	55.56	72.73	60.99
VCC	R@1	8.78	7.27	6.78	11.11	27.27	8.87
	R@5	37.16	36.36	28.81	33.33	45.45	34.75
	R@10	64.86	58.18	**62.71**	55.56	**72.73**	63.12
iReason	R@1	9.27	8.09	7.91	**12.72**	**28.89**	9.21
	R@5	38.71	36.36	29.92	**34.73**	45.75	35.87
	R@10	65.12	58.52	**62.71**	55.56	**72.73**	63.51
OARNet	R@1	**20.95**	14.55	11.86	11.11	9.09	17.38
	R@5	56.76	49.09	37.29	33.33	45.45	50.00
	R@10	75.68	74.55	59.32	55.56	**72.73**	71.28
SEIN (Ours)	R@1	19.59	**16.36**	**16.95**	11.11	27.27	**18.09**
	R@5	**58.78**	**56.36**	**38.98**	33.33	**54.55**	**53.19**
	R@10	**77.03**	**78.18**	61.02	**77.78**	63.64	**73.40**

Table 2: Overall performance compared to the state-of-the-art methods on the test set. The best results are denoted in bold.

Method	R@1	R@5	R@10
w/o SG	16.35	51.22	71.63
w/o OTA	17.21	51.87	71.92
w/o CMF	17.62	52.11	72.24
w/o APS	17.12	52.09	71.88
SEIN	**18.09**	**53.19**	**73.40**

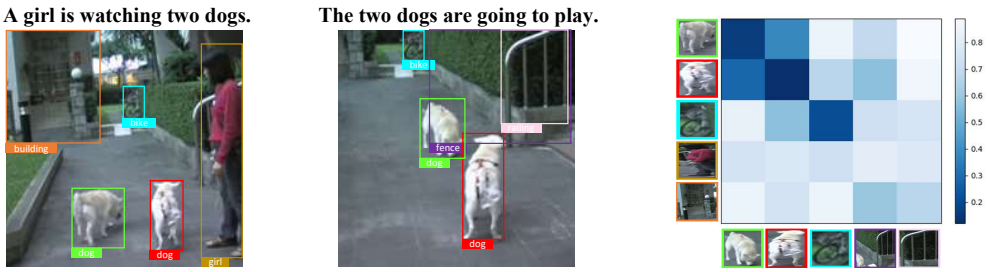
Table 3: Experimental results of ablation study. The best results are denoted in bold.

Main Results

The main experimental results of our method and baselines are reported in Table 2. We can observe that: (1) The proposed method achieves the best performance in terms of R@1, R@5, and R@10 compared to the baseline methods, which suggests the effectiveness of the SEIN framework in addressing this task. Besides, SEIN consistently exhibits excellent performance improvement across different daily life categories. (2) SEIN demonstrates a significant performance gain over the BERT baseline, which indicates that incorporating visual modality can provide valuable information for multimodal event causality reasoning and help rectify certain non-commonsense errors. (3) Compared to VCC and iReason, our method performs far better than them on R@1, R@5, and R@10. The reason behind this improvement may be that VCC and iReason regard the object context representations as features instead of exploiting rich visual features. While our method can make full use of visual and textual features to recognize event causalities more effectively. (4) Our method surpasses OARNet by a substantial margin on R@1, R@5, and R@10. We attribute this to the fact that OARNet primarily uses the co-occurrence of the objects in two images as visual features to identify the causal relation, disregarding the relations between objects in the image and the changes in visual information of the objects. In contrast, SEIN can leverage the structural and semantic information of the image and capture the object interactions across the image pair, thus achieving better performance.

Analysis and Discussion

Ablation Study. To verify the contributions of each component, we conduct ablation studies by comparing SEIN with the proposed variant methods. As illustrated in Table 3, we can find that: (1) After removing scene graph (w/o SG), the model performance drops significantly. The performance gap indicates the importance of the scene graph in modeling the objects and relations between paired objects, which could provide valuable structural and semantic knowledge for each image and facilitate event causality reasoning. (2) After removing optimal transport-based alignment (w/o OTA), the model performance becomes worse. This illustrates that the optimal transport-based alignment strategy can capture the interactions of objects across the image pair, which is beneficial for recognizing the changes in visual information and mining implicit causal clues. (3) After removing the cross-modal fusion module (w/o CMF), the model also suffers from performance decay. This result demonstrates that the multi-head attention mechanism is effective for capturing cross-modal interactions and fusing textual and visual modalities, enabling enhanced reasoning and prediction. (4) After removing the adaptive prediction strategy (w/o APS), which means the causality score is predicted by an average operation, the model performance decreases. This performance gap illustrates that the adaptive prediction strategy can balance the influence of textual and multimodal reasoning for causality prediction, especially in the case of single outcome prediction errors.



(a) The input events and object detection results

(b) Transportation cost matrix

Figure 3: Visualization of a typical instance.

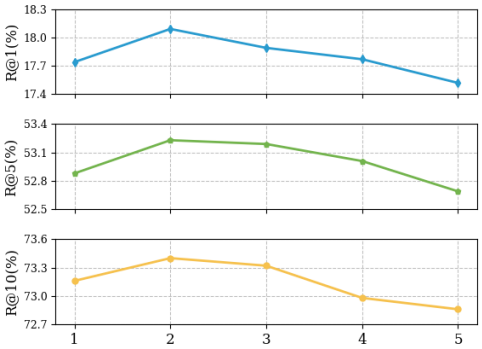


Figure 4: Experimental results under different number of GCN layers.

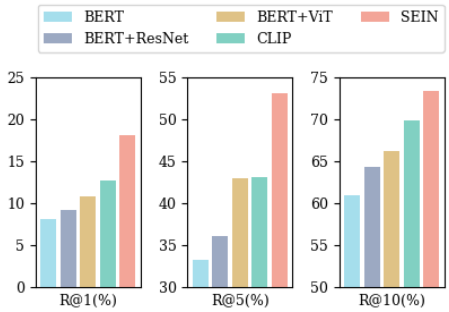


Figure 5: Experimental results of using different pre-trained models.

Effect of the Number of GCN Layers. To investigate the effect of GCN layers, we conduct experiments with the number of GCN layers ranging from 1 to 5. The model performance on R@1, R@5, and R@10 is plotted in Fig. 4. The observations drawn from the results are as follows: (1) SEIN produces the best performance when using two layers on R@1, R@5, and R@10. Therefore, we argue that adopting two GCN layers is most effective in modeling the objects and relations between paired objects to obtain a sufficient understanding of the image. (2) The model performance drops rapidly when the number of GCN layers becomes too large. This illustrates that increasing the number of GCN layers beyond a certain point does not contribute to improving the performance of multimodal event causality reasoning.

Generalization. To prove the generalization ability of the SEIN framework, we leverage different pre-trained models to encode text and image modalities for comparison. The results are presented in Fig. 5. The following observations can be made: (1) SEIN consistently yields the best performance among different methods, indicating the effectiveness and generalization ability of the proposed method. This also suggests that leveraging the structural information of the image and interactions of objects across the image pair enables the model to uncover implicit causal clues, thus boosting reasoning performance. (2) The methods that incorporate visual information generally perform better than BERT, which indicates that the inclusion of global visual features can enhance the model's understanding of multimodal daily events.

Visualization. We present the visualization of a typical instance to demonstrate the object interactions across the image pair. As shown in Fig. 3(a), we adopt the pre-trained Faster R-CNN (Ren et al. 2015) to obtain object information from each image. After training the SEIN framework, the cost matrix from the optimal transport-based alignment strategy is illustrated in Fig. 3(b). We can find that the same kind of objects exhibit relatively lower transportation costs. Additionally, the cost associated with the same object is lower compared to different objects across the image pair. This indicates that using the cost matrix to guide the combination of object features is reasonable and effective.

Conclusion

In this paper, we propose a SEIN framework to tackle the multimodal event causality reasoning task. The proposed method exploits GCN to model the objects and relations from scene graph structure, allowing for a sufficient visual understanding of the image. Then an optimal transport-based alignment approach is designed to capture changes in visual information between the image pair and facilitate causality reasoning. Besides, SEIN adopts a cross-modal fusion module to combine textual and visual features, and introduces an adaptive prediction strategy for better inference. Experimental results illustrate that SEIN achieves state-of-the-art performance on the Vis-Causal dataset.

References

Cao, P.; Zuo, X.; Chen, Y.; Liu, K.; Zhao, J.; Chen, Y.; and Peng, W. 2021. Knowledge-Enriched Event Causality Identification via Latent Structure Induction Networks. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 4862–4872. Association for Computational Linguistics.

Chadha, A.; and Jain, V. 2021. iReason: Multimodal Commonsense Reasoning using Videos and Natural Language with Interpretability. *CoRR*, abs/2107.10300.

Cong, Y.; Liao, W.; Ackermann, H.; Rosenhahn, B.; and Yang, M. Y. 2021. Spatial-Temporal Transformer for Dynamic Scene Graph Generation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 16352–16362. IEEE.

Cuturi, M. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In Burges, C. J. C.; Bottou, L.; Ghahramani, Z.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, 2292–2300.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Hildebrandt, M.; Li, H.; Koner, R.; Tresp, V.; and Günnemann, S. 2020. Scene Graph Reasoning for Visual Question Answering. *CoRR*, abs/2007.01072.

Iashin, V.; and Rahtu, E. 2020. Multi-modal Dense Video Captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, 4117–4126. Computer Vision Foundation / IEEE.

Ji, J.; Krishna, R.; Fei-Fei, L.; and Niebles, J. C. 2020. Action Genome: Actions As Compositions of Spatio-Temporal Scene Graphs. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 10233–10244. Computer Vision Foundation / IEEE.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.*, 123(1): 32–73.

Li, M.; Xu, R.; Wang, S.; Zhou, L.; Lin, X.; Zhu, C.; Zeng, M.; Ji, H.; and Chang, S. 2022. CLIP-Event: Connecting Text and Images with Event Structures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 16399–16408. IEEE.

Li, R.; Zhang, S.; and He, X. 2022. SGTR: End-to-end Scene Graph Generation with Transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 19464–19474. IEEE.

Liu, J.; Chen, Y.; and Zhao, J. 2020. Knowledge Enhanced Event Causality Identification with Mention Masking Generalizations. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, 3608–3614. ijcai.org.

Liu, J.; Zhang, Z.; Guo, Z.; Jin, L.; Li, X.; Wei, K.; and Sun, X. 2023a. KEPT: Knowledge Enhanced Prompt Tuning for event causality identification. *Knowl. Based Syst.*, 259: 110064.

Liu, J.; Zhang, Z.; Wei, K.; Guo, Z.; Sun, X.; Jin, L.; and Li, X. 2023b. Event Causality Extraction via Implicit Cause-Effect Interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6792–6804.

Lu, Y.; Rai, H.; Chang, J.; Knyazev, B.; Yu, G. W.; Shekhar, S.; Taylor, G. W.; and Volkovs, M. 2021. Context-aware Scene Graph Generation with Seq2Seq Transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 15911–15921. IEEE.

Ma, B.; and Tong, C. 2022. Joint Visual Perception and Linguistic Commonsense for Daily Events Causality Reasoning. In *IEEE International Conference on Multimedia and Expo, ICME 2022, Taipei, Taiwan, July 18-22, 2022*, 1–6. IEEE.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global Vectors for Word Representation. In Moschitti, A.; Pang, B.; and Daelemans, W., eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 1532–1543. ACL.

Rajani, N. F.; McCann, B.; Xiong, C.; and Socher, R. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In Korhonen, A.; Traum, D. R.; and Màrquez, L., eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 4932–4942. Association for Computational Linguistics.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Cortes, C.; Lawrence, N. D.; Lee, D. D.; Sugiyama, M.; and Garnett, R., eds., *Advances in*

Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, 91–99.

Sap, M.; Bras, R. L.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 3027–3035. AAAI Press.

Shen, S.; Zhou, H.; Wu, T.; and Qi, G. 2022. Event Causality Identification via Derivative Prompt Joint Learning. In Calzolari, N.; Huang, C.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.; Ryu, P.; Chen, H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S., eds., *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, 2288–2299. International Committee on Computational Linguistics.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased Scene Graph Generation From Biased Training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 3713–3722. Computer Vision Foundation / IEEE.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

Wang, J.; Yang, Y.; Liu, K.; Zhu, Z.; and Liu, X. 2022a. M3S: Scene graph driven multi-granularity multi-task learning for multi-modal NER. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 111–120.

Wang, S.; Wei, K.; Zhang, H.; Li, Y.; and Wu, W. 2022b. Let Me Check the Examples: Enhancing Demonstration Learning via Explicit Imitation. *arXiv preprint arXiv:2209.00455*.

Wang, W.; Wang, R.; Shan, S.; and Chen, X. 2019. Exploring Context and Visual Pattern of Relationship for Scene Graph Generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 8188–8197. Computer Vision Foundation / IEEE.

Wang, Z.; You, H.; Li, L. H.; Zareian, A.; Park, S.; Liang, Y.; Chang, K.-W.; and Chang, S.-F. 2022c. SGEITL: Scene graph enhanced image-text learning for visual commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 5914–5922.

Wei, K.; Sun, X.; Zhang, Z.; Jin, L.; Zhang, J.; Lv, J.; and Guo, Z. 2022. Implicit Event Argument Extraction With Argument-Argument Relational Knowledge. *IEEE Transactions on Knowledge and Data Engineering*.

Wei, K.; Sun, X.; Zhang, Z.; Zhang, J.; Zhi, G.; and Jin, L. 2021. Trigger is not sufficient: Exploiting frame-aware knowledge for implicit event argument extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4672–4682.

Wei, K.; Yang, Y.; Jin, L.; Sun, X.; Zhang, Z.; Zhang, J.; Li, X.; Zhang, L.; Liu, J.; and Zhi, G. 2023. Guide the Many-to-One Assignment: Open Information Extraction via IoU-aware Optimal Transport. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4971–4984.

Wu, B.; Cheng, R.; Zhang, P.; Gao, T.; Gonzalez, J. E.; and Vajda, P. 2022. Data Efficient Language-Supervised Zero-Shot Recognition with Optimal Transport Distillation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Xie, Y.; Wang, X.; Wang, R.; and Zha, H. 2019. A Fast Proximal Point Method for Computing Exact Wasserstein Distance. In Globerson, A.; and Silva, R., eds., *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, volume 115 of *Proceedings of Machine Learning Research*, 433–453. AUAI Press.

Xu, B.; Zeng, Z.; Lian, C.; and Ding, Z. 2022. Few-Shot Domain Adaptation via Mixup Optimal Transport. *IEEE Trans. Image Process.*, 31: 2518–2528.

Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene Graph Generation by Iterative Message Passing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 3097–3106. IEEE Computer Society.

Yoon, S.; Kang, W.; Jeon, S.; Lee, S.; Han, C.; Park, J.; and Kim, E. 2021. Image-to-Image Retrieval by Learning Similarity between Scene Graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 10718–10726. AAAI Press.

Zareian, A.; Karaman, S.; and Chang, S. 2020. Weakly Supervised Visual Semantic Parsing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 3733–3742. Computer Vision Foundation / IEEE.

Zhang, H.; Huo, Y.; Zhao, X.; Song, Y.; and Roth, D. 2021. Learning Contextual Causality Between Daily Events From Time-Consecutive Images. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, 1752–1755. Computer Vision Foundation / IEEE.