Visual Relationship Transformation

Xiaoyu Xu[1], Jiayan Qiu[2]⋆, Baosheng Yu[3], and Zhou Wang[1]

[1] University of Waterloo, Canada
[2] University of Leicester, UK
[3] University of Sydney, Australia
`x423xu@uwaterloo.ca, jiayan.qiu.1991@outlook.com, baosheng.yu@sydney.edu.au, zhou.wang@uwaterloo.ca`

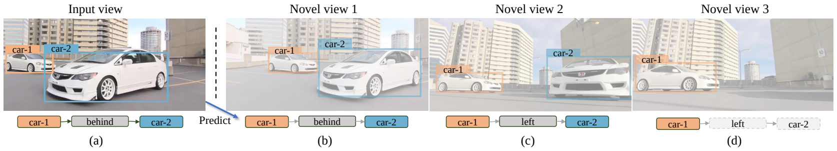


Fig. 1: Visual Relationship Transformation. Given an observed view image (a), we aim to predict the relationships between objects (*e.g.*, the one between `car-1` and `car-2`) in novel views (b-d) without requiring their corresponding view images. Compared with the relationship observed in (a), the corresponding ones in novel views (b-d) are typically view-dependent and nondeterministic, which might be preserved, altered, or even become invisible. The faded images and relationships denote unseen and invisible, respectively.

Abstract. What will be the relationships between objects in a novel view? We strive to answer this question by investigating a new visual cognition task, termed visual relationship transformation or VRT. Unlike prior visual relationship detection task that works on visible view images, VRT aims to predict the relationships in unseen novel views from a single observed source view. Towards solving VRT, we propose an end-to-end deep approach that, given an observed view image and inter-view transformations, learns to predict the relationships in novel views. Specifically, we introduce an equivariant graph neural network to predict the relationships between objects in novel views, which is achieved by enforcing the transformation equivariance of the learned relationship representations. Simultaneously, a relationship presentness mask is learned for pruning the invisible ones, thus enabling the visible relationship prediction in novel views. To this end, VRT provides supplementary cues for accomplishing novel-view-related tasks, such as visual grounding (VG), novel view synthesis (NVS), and pedestrian intention estimation (PIE). In the experiments, adopting VRT as a plug-in module results in considerable performance improvements in VG, NVS, and PIE across all datasets.

1 Introduction

Looking at the visual relationship `(car-1, behind, car-2)` in Fig. 1(a) and asking to predict its corresponding ones in the novel views shown in Fig. 1(b-d),

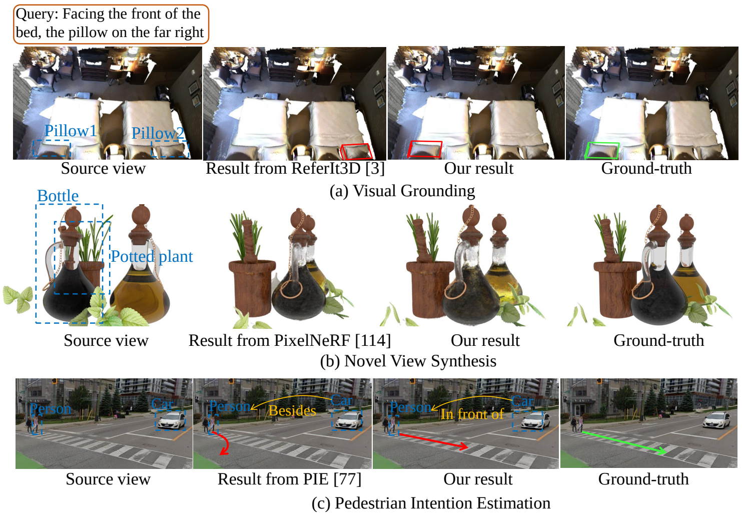
⋆ Corresponding author

it can be found that the transformation of visual relationships is view-dependent and non-deterministic. The transformed relationships might be preserved, altered, or even become invisible in novel views. Therefore, given a single observed view image and the cross-view transformations, can we learn to adaptively predict the transformed relationships and their presentness in novel views? We term this task, which has never been explored in prior works, as *visual relationship transformation* or VRT. The VRT plays a vital role in scenarios where a multiple-view understanding of a scene is demanded, particularly when the acquisition costs are expensive. For instance, when a robot is tasked with locating a referred object, it is necessary to understand the relationships between objects in the scene from various perspectives. Nonetheless, observing from every conceivable angle is expensive or impractical. So predicting the relationships in unseen views given observed views is a tractable and economical solution.

Comparing with numerous efforts in visual relationship detection or VRD [9, 34, 39, 40, 55, 56, 70, 73, 76, 89, 93, 95, 97], VRT stands out by explicitly predicting the corresponding relationships in unseen views. In contrast, a vanilla solution, which first performs NVS and then conducts VRD, not only requires huge computations but also fails to maintain the correspondence between relationships from the source and target views. Furthermore, VRT comes with a large amount of supplementary cues, such as structural understanding, that help novel-view-related tasks. For example, the visual grounding task [2] shown in Fig. 2(a), which asks to identify the '`pillow on the far right`' from the unseen view of '`the front of the bed`', requires VRT for predicting the transformed relationships in the novel view, thus can perform accurate alignment between the textual and visual semantics. Moreover, the novel-view-synthesis task [90], shown in Fig. 2(b), requires VRT to accurately predict the relationship transformation to the novel view, thus ensuring the correct relationship between the `potted plant` and the `bottle` in the synthesized image. Furthermore, the pedestrian intention estimation task [61], shown in Fig. 2(c), requires VRT to predict the relationships in the pedestrian view from the given camera view image, thus being able to estimate the pedestrian intention accurately.

While VRT can provide an informative understanding of novel views, learning to solve it remains a challenge. Firstly, labeling relationships in multi-view datasets is expensive, making large-scale training data unavailable. Secondly, in the novel view, the prediction of relationships and their corresponding presentness should be learned separately. As depicted in Fig. 1(d), the relationship should be predicted as `left` even though it is invisible in the novel view, where invisibility differs from unrelatedness. Thirdly, the predicted scene graph, constructed by the predicted relationships, should be structurally consistent with the ground-truth scene graph in the novel view, thus allowing the learning of global structure information.

Towards accomplishing VRT, we devise a novel approach to perform relationship transformation and tackle the aforementioned challenges during training. Specifically, given a single observed view image, we first detect the objects with an off-the-shelf detection algorithm. Each detected object is modeled as a node in



(a) Visual Grounding

(b) Novel View Synthesis

(c) Pedestrian Intention Estimation

Fig. 2: Illustration of VRT implementations on Visual Grounding, Novel-View Synthesis, and Pedestrian Intention Estimation tasks. For each task, the first image denotes the given input, the second one denotes the result from the SOTA method, the third one denotes the result by adopting our VRT as a plug-in module, and the fourth one denotes the ground-truth.

the scene graph. This scene graph, along with the cross-view transformation, is then fed into the VRT module for predicting the transformed relationships, *e.g.* the edges in the scene graph, and their corresponding presentness. Throughout this process, the transformation equivariance of the relationship representations is enforced. To ensure the proper learning of relationship transformation, in the training stage, we first extract the ground-truth relationship features, presentness, and the scene graph from the novel view using a pre-trained visual relationship detection model [86]. Then, a relationship transformation loss is introduced to supervise each relationship prediction by adopting the ground-truth features as supervision, thus mitigating the lack of training data. Meanwhile, a presentness loss is implemented to learn the visibility of relationships in the novel view. Finally, a relationship structure loss is introduced to ensure the global structural information in the novel view can be learned. With these components, the trained approach is capable of performing visual relationship transformation in a structurally consistent manner.

Our contribution includes both a new VRT task and a novel approach designated to address the proposed task. To the best of our knowledge, this represents the first attempt in this direction. To evaluate the recognition accuracy of our VRT, we conducted extensive objective and subjective experiments and our VRT obtained encouraging results. Moreover, by adopting our VRT as a plug-in module, state-of-the-art methods in visual grounding [2], novel-view synthesis [83], few-shot NERF [90], and pedestrian intention estimation [61] have achieved considerable qualitative and quantitative performance improvement on Nr3D [2], Realestate10K [98], DTU [27], and PIE [61] datasets, respectively.

2 Related Work

In this section, we briefly review prior works related to ours, including visual relationship detection, equivariant neural networks, and novel view synthesis.

Visual Relationship Detection. The visual relationship detection methods can be categorized into three types: object-related [17], property-related [23], and activity-related [88]. In this work, we focus on object-related relationships, where early approaches involve the recognition of visual phrases [19,39,41]. With the significant development of deep learning, recent methods demonstrate encouraging performance improvements due to their high representation capability [7,15,20,29,33,40,41,43–45,49,54,93,94,99]. More recently, video relationship detection methods [8,38,46,56,69,70,72,73,76,97] have been proposed to capture temporal dynamics in relationships. However, none of these works has explored predicting visual relationships in unseen views.

Equivariant Neural Networks. The equivariant neural networks are proposed to preserve the transformation equivariance, such as translation and rotation, of the learned representations from the images [11]. Recently, steerable CNNs and 3D-steerable CNNs are proposed for simplifying the equvairant convolutional process [12–14, 81, 82]. Meanwhile, gauge-equivariant networks are proposed to enforce the equivariance with local transformations [10]. More recently, equivariant graph neural networks are designed to process non-grid data [3–5,31], which preserves the appearance equivariance of the point clouds and 3D volumetric data. However, all existing methods ignore the equivariant message passing manner on edge processing of the graph data, thus cannot be directly implemented on VRT task.

Novel View Synthesis. Neural rendering methods can be categorized into four types: point-based [18,25,42,47,58,60,63,66,67,75,79,87,92], mesh-based [30, 48], surface-based [6, 21, 53, 71], and volumetric-based [24, 26, 36, 57, 96]. During the rendering process, deterministic rendering kernels [100, 101] and global semantic understandings [32, 35, 59, 83] are adopted to improve performance. Recently, neural radiance field (NeRF) methods [50–52, 78] have been proposed to formulate scenes as learned neural networks, enabling novel view synthesis. More recently, generalizable NeRF [62, 77, 84] methods for complex scenes have been introduced. However, none of these approaches attempts to utilize cues from visual relationships in the synthesis process.

3 Method

In this section, we present the main VRT learning scheme. As illustrated in Fig. 3, our approach consists of three stages. In **Stage 1**, we initiate the process by detecting objects in the source view to construct a scene graph. This graph, coupled with the cross-view transformation, is then input into the VRT module for predicting transformed relationships and determining their presentness in the target view. In **Stage 2**, we search for matches between the predicted relationships, presentness, and the scene graph constructed accordingly, and their corresponding ground truth extracted from the target view. Finally, **Stage 3** introduces specific loss components: the relationship transformation loss for single

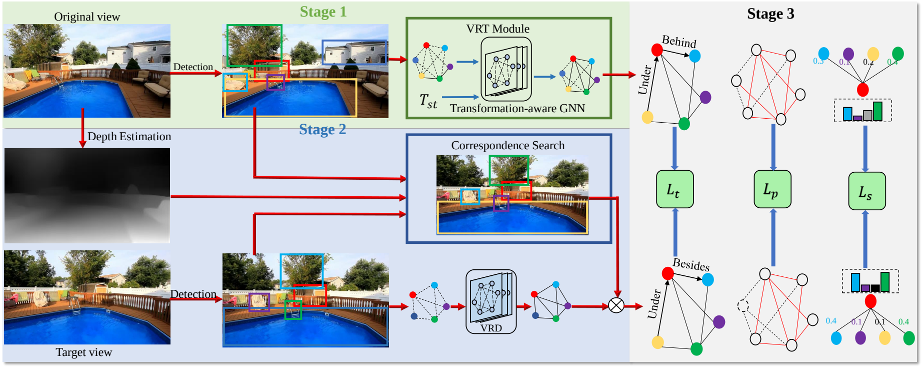


Fig. 3: The illustration of the proposed approach. The T_{st} denotes the transformation matrix from the observed source view to the target view. $\bigotimes$ denotes the re-indexing and alignment operation. Note that, the relationship values and presentness are predicted by the VRT module, and the structure information is constructed by the predicted relationship and presentness.

relationship prediction, the presentness loss for predicting relationship visibility, and the relationship structure loss to preserve essential structural information.

3.1 Stage 1: VRT

In real-world scenarios, the observation of a scene from different views involves cross-view transformations, encompassing both rotation and translation, thus the observed relationships among objects in the scene should adhere to the spatial transformation. For example, in Fig. 1(b) and (c), the relationship between two cars changes from `behind` to `left` when observed from the side to the front. These two relationships are linked through the view transformation and conditioned on the scene. Therefore, we aim to predict the inter-object relationships in the target view while also keeping track of the changes in these relationships from the source to the target view. Motivated by the concept of equivariant neural networks, which preserves relationship correspondences over view transformations, an equivariant graph neural network is designed for learning transformation-equivariance representations of the relationships between objects.

Scene Graph Construction. Given an image from the original or source view, we initially employ an off-the-shelf object detector [64] to detect N objects and extract various pieces of information. This includes visual embeddings $\mathcal{X} = \{x_1, \ldots, x_N\}$, positional information in the form of bounding boxes $\mathcal{B} = \{b_1, \ldots, b_N\}$, semantic embeddings $\mathcal{W} = \{w_1, \ldots, w_N\}$ (i.e., word vector representing categories), and union region representations $\mathcal{U} = \{u_{1,1}, \ldots, u_{N,N}\}$. Subsequently, we construct a scene graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ in the following manner. The node feature $v_i \in \mathcal{V}$ and the edge (or relationship) feature $e_{i,j} \in \mathcal{E}$ are defined as

$$\begin{aligned} v_i &= \phi_v(x_i \oplus \mathtt{PE}(b_i) \oplus w_i), \\ e_{i,j} &= \phi_e(u_{i,j} \oplus v_i \oplus v_j), \end{aligned} \tag{1}$$

where $\oplus$ denotes the concatenation operation, $\texttt{PE}(\cdot)$ indicates the positional encoding operation [22]. ϕ_v and ϕ_e denote the equivariant layers [16], which firstly lifts the the size of the input from $N \times 1$ to $N \times 3$, and then perform representation learning with the equivariance-constraint:

$$\phi(\mathcal{T}_g'(x)) = \mathcal{T}_g(\phi(x)), \tag{2}$$

where $\mathcal{T}_g : X \to X$ is a set of transformations for the abstract group g.

Equivariant Message Passing. After the scene graph is constructed, we design an Equivariant Message Passing (EMP) module to model the correlations among nodes and relationships while maintaining the equivariance property. Starting with the initial node and relationship features v_i^0 and $e_{i,j}^0$ from the previously constructed scene graph, the node message passing at step $l+1$ is formulated as follows:

$$\begin{gathered} v_i^{l+1} = v_i^l + \sigma(M_{i\cdot}^v + M_{\cdot i}^v), \qquad (3) \\ M_{i\cdot}^v = \sum_{j \in \eta_i} \alpha_{ij} \phi_1(e_{ij}^l), M_{\cdot i}^v = \sum_{j \in \eta_i} \alpha_{ji} \phi_1(e_{ji}^l) \qquad (4) \\ \alpha_{ij} = \frac{\exp(\sigma(\phi_2[v_i \oplus v_j]))}{\sum_{k \in \eta_i} \exp(\sigma(\phi_2[v_i \oplus v_k]))}, \end{gathered}$$

where σ refers to Vector-Neuron ReLU [16] and ϕ_1, ϕ_2 represent different Vector-Neuron MLPs (VN-MLPs) [16]. The term $M_{i\cdot}$ denotes the weighted averaged messages from the relationships between the i-th object and its neighbor objects η_i, while the $M_{\cdot i}$ denotes message from neighbors to the i-th object. Additionally, the formulation for relationship message passing is as follows:

$$\begin{gathered} e_{i,j}^{l+1} = e_{i,j}^l + \sigma(M_{i\cdot}^e + M_{j\cdot}^e), \qquad (5) \\ M_{i\cdot}^e = \sum_{k \in \eta_i} \beta_{ik}(\phi_3([v_k \oplus \phi_4(b_i - b_k)])), \\ \beta_{ik} = \frac{\exp(\sigma(\phi_5[v_i \oplus v_k \oplus \phi_4(b_i - b_k)]))}{\sum_{l \in \eta_i} \exp(\sigma(\phi_5[v_i \oplus v_l \oplus \phi_4(b_i - b_l)]))}, \end{gathered}$$

where the $M_{i\cdot}^e$ and $M_{j\cdot}^e$ denote the aggregated message from the i-th and j-th objects to the relationship $e_{i,j}$. β_{ik} measures the contribution of message $M_{i,k}$ to the relationship $e_{i,j}$. ϕ_3 and ϕ_5 denote VN-MLPs, and ϕ_4 denots a VN-MLP with lifting operation [16]. The process of equivariant relationship message passing integrates relative position information into the relationship embedding while preserving equivariance, as the relative position is inherently equivariant to transformations such as rotation and translation.

Relationship Transformation. After the equivariant message passing, we predict the inter-object relationships in the target view. Let $\mathcal{E}_s = \{e_{i,j}\}$ denote the relationship representations from the source view V_s, for any target view V_t, we then predict its relationship representations $\hat{\mathcal{E}}_t = \{\hat{e}_{i,j}\}$ as follows:

$$\hat{\mathcal{E}}_t = \phi(T_{st} \cdot \mathcal{E}_s)^\top, \tag{6}$$

where ϕ represents an MLP layer, and T_{st} indicates the transformation matrix from the source view V_s to the target view V_t. In addition to the predicted relationship representations $\hat{\mathcal{E}}_t$, we also predict their presentness as follows:

$$\hat{p}_{i,j} = \texttt{sigmoid}\left(\texttt{MLP}\left(\left[v_i^l, v_j^l, e_{i,j}^l\right]\right)\right), \tag{7}$$

where l refers to the last message passing step, $\hat{p}_{i,j}$ indicates the probability of the relationship between node v_i and v_j being present in the target view. During inference, the presentness is determined as $\mathbf{1}\{\hat{p}_{i,j} \geq \gamma\}$, where γ is a threshold. In other words, it signifies that the relationship between the i-th and the j-th object is visible in the target view.

3.2 Stage 2: Correspondence Search

As described above, the VRT predicts the relationships and their corresponding presentness in the target view V_t based on the source view V_s and the inter-view transformation matrix T_{st} between the source and target views. To facilitate the training of VRT, it is necessary to align the predicted scene graph $\hat{\mathcal{G}}_t$ with the ground truth scene graph $\mathcal{G}_t$ in the target view V_t. To accomplish this alignment, we initially transform all object bounding boxes from the source view V_s to the target view V_t using the transformation matrix T_{st}. For each pixel (x, y) within a bounding box, we determine the transformed pixel coordinates $(\hat{x}, \hat{y})$ as follows:

$$[\hat{x}, \hat{y}, 1]^\top = K \cdot T_{st} \cdot D(x, y) \cdot K^{-1}[x, y, 1]^\top, \tag{8}$$

where $D(x, y)$ denotes the depth at pixel (x, y), obtained using a pre-trained depth estimation algorithm [65], and K represents the camera intrinsic matrix (i.e., both the source and target views use the same camera). With this setup, we can evaluate the overlap between each transformed bounding box and all bounding boxes in the target view. Specifically, if the transformed bounding box $\hat{b}_i$ has the maximum overlap with a target bounding box b_j among all bounding boxes in the target view, and $\hat{b}_i$ and b_j share the same object category label, we then identify a pair of nodes matched between the transformed scene graph and the target scene graph. For nodes and edges transformed from the source view that are invisible in the target view (i.e., fail to match with any others), we assign their ground truth presentness labels as 0, otherwise as 1.

3.3 Stage 3: Optimization

Through the identification of correspondences between the scene graph transformed from the source view and the ground truth scene graph in the target view, we train the proposed VRT model by aligning specific components of the two scene graphs using three different loss functions, as outlined below.

For every pair of the predicted (or transformed) relationship $\hat{e}_{i,j}$ and its corresponding ground truth relationship $e_{i,j}^*$, we define the **relationship transformation loss** as:

$$\mathcal{L}_t = \frac{1}{N_m}\sum_{i,j} \|\hat{e}_{i,j} - e_{i,j}^*\|_2^2, \tag{9}$$

where N_m denotes the number of matched relationship pairs, $e^*_{i,j}$ comes from the pre-trained Graph-RCNN [86] by taking the target view as input. For the prediction of relationship presentness $\hat{p}_{i,j}$, we define the **relationship presentness loss** as:

$$\mathcal{L}_p = \frac{1}{N_p} \sum_{i,j} \|\hat{p}_{i,j} - p_{i,j}\|_2^2, \tag{10}$$

where N_p represents the total number of edges (or relationships) in the scene graph constructed from the source view, and $p_{i,j}$ indicates the ground truth presentness label derived from the target view. In addition to relationship prediction, it is crucial to appropriately preserve the topology structure constructed by them. Inspired by [87], we characterize the topology structure of each edge $e_{i,j}$ through a set of its neighboring edges $\mathcal{N}(e_{i,j})$ that share one of the same nodes. Specifically, the information about the neighbor structure of edge $e_{i,j}$ is defined as a normalized distribution vector $\texttt{NS}_{i,j} \in \mathbb{R}^{|\mathcal{N}(e_{i,j})|}$. For each $e_{m,n} \in \mathcal{N}(e_{i,j})$, its corresponding element in $\texttt{NS}_{i,j}$ can be evaluated as:

$$\texttt{NS}_{i,j} = \frac{\exp\{||e_{i,j} - e_{m,n}||_2^2\}}{\sum_{e_{k,l} \in \mathcal{N}(e_{i,j})} \exp\{||e_{i,j} - e_{k,l}||_2^2\}}. \tag{11}$$

With the above definition of the neighbor structure, we then define the **relationship structure loss** as follows:

$$\mathcal{L}_s = \frac{1}{N_r} \sum_{(i,j)} \mathcal{D}_{\texttt{KL}}(\hat{\texttt{NS}}_{i,j} || \texttt{NS}_{i,j}), \tag{12}$$

where $\mathcal{D}_{\texttt{KL}}$ denotes the Kullback-Leibler divergence, and $\hat{\texttt{NS}}_{i,j}$ and $\texttt{NS}_{i,j}$ represent the predicted and ground truth neighbor structure information, respectively. Finally, the overall loss function for training the proposed VRT is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_t + \lambda_2 \mathcal{L}_p + \lambda_3 \mathcal{L}_s, \tag{13}$$

where λ_1, λ_2, and λ_3 denote the weights used to balance different components.

4 Experiments

In this section, we present our experimental setups and results. Since no existing work performs exactly the same task as we do here, we mainly focus on showcasing the effectiveness of the proposed approach in two main aspects. Firstly, we compare the relationship prediction performance of our VRT approach with scene graph generation methods relying on visible target views. Additionally, through the adoption of VRT, we demonstrate its enhancements in various related tasks, such as novel view synthesis (leveraging Synsin [83] and PixelNeRF [91]), pedestrian intention prediction (based on [61]), and visual grounding (based on ReferIt3D [2]). It's important to note that our objective is to showcase the potential of learning visual relationship transformation to enhance multi-view understanding, rather than attempting to outperform all state-of-the-art methods.

4.1 Datasets and Implementation Details

We present the datasets used in our experiments as follows: For relationship prediction experiments, we employ the Realestate10K dataset [98] and assess generalizability using the VidVRD dataset [28]. Novel view synthesis experiments utilize both the Realestate10K dataset [98] and the DTU dataset [1]. Pedestrian intention estimation and trajectory prediction experiments rely on the PIE dataset [61]. Visual grounding experiments involve the Nr3D dataset [2]. Below, we provide a brief introduction to each dataset.

Realestate10K [98]. It contains 80K video clips from both indoor and outdoor scenes, comprising more than 10 million frames annotated with camera intrinsic parameters and camera poses obtained from SfM. Following [83], we use 57K, 14K, and 7K scenes for training, validation, and testing, respectively. Notably, all scenes in the test set are unseen. We sample views by first selecting an original view and then choosing the target view with a maximum of 70 frames apart. Subsequently, we select frame pairs with an angle change of $\geq 5^\circ$ and a position change of ≥ 0.15. Finally, we obtained 16K, 4K, and 2K frame pairs for training, validation, and testing, respectively. In the novel view synthesis experiment, all selections are the same except for the inter-frame distance. The target views are chosen with a maximum of 30 frames apart from the original view, and we report the results on 2K samples from [83].

VidVRD [28]. It contains 1000 videos with 4835 annotated relationship instances spanning 132 categories. The training and validation sections consist of 800 and 200 videos, respectively. Due to the limited number of relationship annotations in VidVRD, we select a total of 500 image pairs from the validation videos to test the generalizability of VRT. The selected image pairs adhere to the same criteria for angle and position changes as in Realestate10K. The transformation matrices are generated using COLMAP [68].

DTU MVS [1].]. For each object, it contains 49 or 65 images, with camera information, baseline, and structured light ground-truth. The presented objects pose a challenge as their appearance changes in different viewpoints due to specularities. We adopt this dataset for novel view synthesis, following the training settings in [91].

PIE [61]. It consists of over 6 hours of video footage capturing pedestrians in various types of crosswalks, recorded by an onboard camera. It includes both ground-truth pedestrian intentions and trajectories, along with bounding boxes for traffic objects, pedestrian attributes, and camera parameters. There are 1,842 pedestrian samples and 293K annotated frames. Following [61], we use 150K, 30K, and 113K frames for train, validation, and test sets, respectively.

Nr3D [2]. It annotates 707 indoor 3D scenes with 45,503 human utterances. Follow the setting of ReferIt3D [2], we use 30K, 2K, and 8K image-query pairs for training, validation, and testing, respectively. Furthermore, during the testing, we report the results on both view-dependent and view-independent image-query pairs.

Implementation Details. The proposed approach is implemented with Pytorch on four NVIDIA V100 GPUs. In the training process, the batch size is set to 16. Adam optimizer is adopted with a fixed learning rate of 0.0001. The loss weights λ_1, λ_2, and λ_3 are set to be 0.5, 0.5, and 0.1. During inference, the threshold γ is set to be 0.5.

4.2 Visual Relationship Transformation

		SGCls		PredCls	
Method	Input View	R@50	R@100	R@50	R@100
VRT	Original	30.3	34.7	33.6	37.8
Synsin [83]+GRCNN [86]	Original	23.7	24.1	25.3	27.7
Synsin [83]+TDE [74]	Original	23.8	24.3	25.6	27.8
Synsin [83]+BGNN [37]	Original	24.0	24.4	25.2	28.0
GRCNN [86]	Target	31.0	34.9	34.0	38.2

Table 1: Results evaluated with Recall@50 and Recall@100 on tasks: relationship classification (PredCls) and scene graph classification (SGCls).

We conduct both objective and subjective evaluations on hand-labeled Realestate10K [98] datasets. Then, we conduct the generalizability test on the VidVRD dataset. Specifically, we classify transformed relationships using a pre-trained classifier [86]. Only the relationships predicted to be visible in the target view are considered for evaluation.

Objective Evaluation. Following the relationship categories from the Visual Genome dataset, on which the VRD is pre-trained, we manually label 500 image pairs from the test set of Realestate10K, encompassing a total of 10K relationships across 51 categories. These hand-labeled images are used only for testing purposes. Subsequently, we conduct a comparative analysis with 1) the VRD test on the synthesized view initially takes the original view image as input, synthesizes the novel view using Synsin [83], and then detects visual relationships on the synthesized view; 2) Graph-RCNN [86], TDE [74] and BGNN [37] tests are implemented on the target view, which is unavailable for our VRT model; and 3) Our VRT predicts the relationship in the target view based on the original view image. In Table 1, when provided with the same original views as input, the VRT outperforms the VRD on the synthesized view. Furthermore, despite making predictions without the target view, our VRT achieves results comparable to Graph-RCNN [86]. The results also reveal that despite a better VRD model such as TDE [74] being given, the performance on synthesized novel views remains constrained. This underscores that the quality of the synthesized views is a significant bottleneck. Therefore, it proves that directly predicting transformed relationships from the original view is more effective and efficient.

Generalizability Evaluation. We further validate the generalizability of the VRT on VidVRD [28] dataset. Since the relationship categories from VidVRD are different from the pre-trained Graph-RCNN [86], we impose a relationship mapping from the visual-genome [34] annotations to the VidVRD [28] annotations, details of which can be found in supplementary materials. The VRDFormer [97] is tested with the target view as input, while our VRT only takes the original view as input. The results in Tab.2 show the VRT outperforms the VRDFormer [97] which demonstrates the VRT successfully learns multi-view understanding when only the original view is given.

4.3 Ablation Studies

Transformation Sensitivity. Given that the proposed approach is transformation-related, we conduct experiments here to test the sensitivity of our VRT to transformations. Specifically, we split the testing frame pairs into three sets: 1) T_1 with angles $\leq 10°$ & translations ≤ 0.20; 2) T_2 with angles $\leq 15°$ & translations ≤ 0.25; and 3) T_3 with angles $> 15°$ *or* translations > 0.25. In Tab. 3, we see that VRT performs well at different transformation levels, demonstrating that VRT is transformation-insensitive.

Method	Input View	R@50	R@100
VRDFormer [97]	Target	18.59	22.37
VRT	Original	21.35	25.41

Table 2: Results compared with VRDFormer [97].

	Recall@50	Recall@100
T_1	31.5	35.8
T_2	30.6	34.2
T_3	29.8	33.6

Table 3: Results of VRT with different transformations.

	Recall@50	Recall@100
T in nodes	29.5	32.4
T in edges	29.4	32.2
T in attention	29.2	31.5
T in MP	30.3	34.7

Table 4: Results of involving the transformation information in different stages of information aggregation.

	Recall@50	Recall@100
VRT	30.3	34.7
VRT w label	22.2	23.6
VRT w/o $\mathcal{L}_p$	27.8	28.7
VRT w/o $\mathcal{L}_s$	25.5	27.0

Table 5: Results of ablation study.

Transformation Awareness. During the information aggregation of VRT, the equivariant GNN can involve the transformation information in different stages. In this experiment, we conduct experiments on the hand labeling data to evaluate the performance of involving a transformation in only the nodes representation, only the edge representation, only the attention calculation, and the whole message-passing process. As shown in Tab. 4, the message-passing one performs the best. This is because the transformation equivariance should be preserved by the node representations, edge representations, and attention, which can only be effectively captured by the message-passing process.

Ablation Study. Here, we show the effectiveness of different loss terms. Specifically, in the first experiment, we replace the predicted and ground-truth representation with one-hot coding, where the predicted one-hot is obtained from `argmax` and the ground-truth one comes from the pre-trained Graph-RCNN [86]. Then, we conduct experiments without either the relationship presentness loss or the relationship structure loss. As shown in Tab. 5, after replacing the representations with one-hot labels, the performance decreases a lot. This is because, the same visual content can be annotated as different relationships, where the one-hot label breaks the similarity between different relationships. Furthermore,

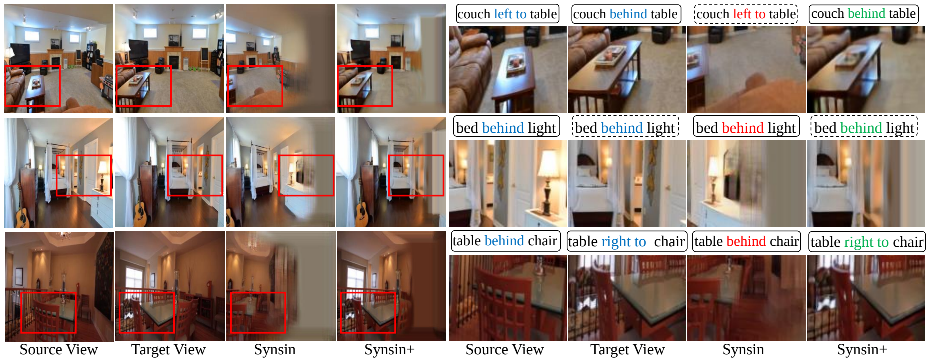


Fig. 4: Visual results from Synsin [83] on Realestate10K [98]. The first two columns show the images from the original view and the target view, the third one shows the synthesized images from Synsin, and the fourth one shows the images from Synsin+ that adopt VRT as a plug-in module. The right four columns present that the synthesized images get better quality when relationships are predicted correctly. The dashed boxes denote invisible relationships in the target views.

after removing the relationship structure loss, the performance decreases dramatically, which shows the importance of preserving the structural information.

4.4 Novel View Synthesis

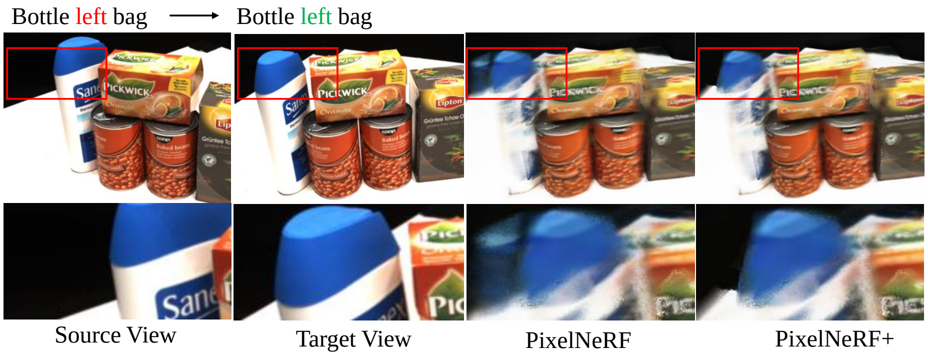
We evaluate the effectiveness of VRT on novel view synthesis tasks. Instead of hard constraining the view synthesis by providing the one-hot relationship prediction, we utilize the relationship features from the pre-trained VRT model in the process of synthesizing, which is achieved by adopting the VRT model as a plug-in module. During which, the relationship features are spatially aligned into the image space by relationship spatialization operation in [85]. Specifically, two state-of-the-art novel view synthesis methods, Synsin [83] and PixelNeRF [90], are used for evaluation.

	PSNR ↑	SSIM [80]↑	Perc Sim ↓
Synsin+	22.78	0.810	1.00
Synsin	22.31	0.740	1.18

Table 6: Results on Realestate10K [98] dataset. Synsin+ denotes enhanced with VRT

Synsin [83]. We evaluate the effectiveness of VRT on Synsin by adopting VRT as a plug-in module to provide relationship features (denoted as Synsin+), concatenating them with the global image features extracted by the sub-module of Synsin. Following [83], the model is retrained on Realestate10K [98]. As shown in Fig. 4, VRT helps preserve the correct visual relationships in view synthesis. The quantitative results are shown in Tab. 6, where VRT brings considerable performance improvement on all metrics.

Upon observation, we note that Synsin may display insensitivity to large transformations, potentially resulting in structural inconsistencies. For example, in the seventh image of the third row in Fig. 4, given large the inter-view



Source View Target View PixelNeRF PixelNeRF+

Fig. 5: Visual results from PixelNeRF [90] on DTU MVS [1].

transformation, Synsin generates the inadequately transformed image, while the structural consistency is enhanced by VRT as shown in the third image, possibly because composing predicted relationships helps preserve the structural information during transformation, thus enhancing the consistency between the synthesized image and the target one. In Tab. 7, we follow the same setting in Sec. 4.3 to evaluate VRT on preserving structural consistency at different transformation scales, where Synsin+ becomes more robust to transformations.

PixelNeRF [90]. To evaluate VRT on PixelNeRF, we adopt it as a plug-in module to provide relationship features, which are then concatenated with the global image features extracted by the sub-module of PixelNeRF. Following the training setting in [90], the model is retrained on DTU MVS [1] dataset. The visual results are shown in Fig. 5, where three views are used for training PixelNeRF and we adopt one of them in VRT to predict the transformed relationships. As shown in Tab. 6, the model with VRT achieves considerable performance improvement on all metrics among all training settings.

		PSNR	SSIM [80]	Perc Sim
T_1	Synsin+	19.98	0.712	1.301
	Synsin	19.83	0.678	1.431
T_2	Synsin+	19.90	0.685	1.210
	Synsin	18.34	0.600	1.285
T_3	Synsin+	19.82	0.732	1.022
	Synsin	17.86	0.584	1.130

Table 7: Quantitative results on different transformation scales.

	PSNR	SSIM [80]	Perc Sim
PixelNeRF(1 v)+	15.91	0.618	0.465
PixelNeRF(1 v)	15.55	0.537	0.535
PixelNeRF(3 v)+	20.28	0.713	0.298
PixelNeRF(3 v)	19.333	0.695	0.387

Table 8: Results on DTU MVS [1]. 1v and 3v denote the number of available views in the training process.

	Acc ↑	F1 ↑	MSE ↓	C-MSE ↓
PIE+	0.8960	0.9446	551.29	513.23
PIE [61]	0.7905	0.8747	559.377	520.53

Table 9: Results on PIE [61].

	Overall	View-dep	View-indep
ReferIt3D [2]	0.356	0.325	0.371
ReferIt3D+	0.455	0.467	0.451

Table 10: Results on Nr3d [2] datasets.

4.5 Pedestrian Intention Estimation

We evaluate the effectiveness of VRT in estimating pedestrian crosswalk intention by adopting it as a plug-in module for [61]. Specifically, instead of utilizing car-view relationships, our VRT provides the feature of relationships in the

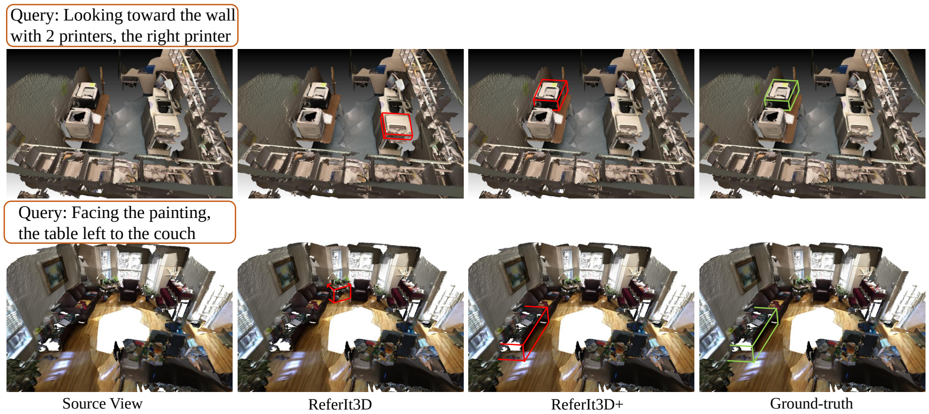

Query: Looking toward the wall with 2 printers, the right printer
Query: Facing the painting, the table left to the couch
Source View
ReferIt3D
ReferIt3D+
Ground-truth

Fig. 6: Visual results from RerferIt3D [2] and RerferIt3D+.

pedestrian view, which is concatenated with the global image features extracted by the CNN module, and then fed into the LSTM module for intention estimation. Since it also benefits the prediction of pedestrian trajectory in a similar way, we thus conduct experiments on both pedestrian intention estimation and trajectory prediction. As shown in Tab. 9, both two tasks gain significant performance improvements. This is because humans perform crosswalks and trajectory planning according to the inter-object relationships in their view, rather than the view of the car.

4.6 Visual Grounding

We evaluate VRT for visual grounding by integrating it with ReferIt3D [2]. As shown in Fig. 6, RerferIt3D+ enhanced by VRT can locate more accurately than ReferIt3D [2]. Furthermore, we train and test ReferIt3D [2] enhanced by VRT, and Tab. 10 shows the improvements on location accuracy, where the improvement on view-dependent queries is larger than the view-independent. This demonstrates that VRT endows ReferIt3D [2] with multi-view understanding.

5 Conclusion

In this paper, we introduced visual relationship transformation (VRT), a novel task aiming to predict the relationships between objects in unseen views, which has never been explored before. Our proposed approach involves training equivariant GNNs to learn VRT, and predict relationships and their presentness in unseen views. To ensure proper prediction and structural consistency of visible relationships, we introduce a relationship transformation loss, a relationship presentness loss, and a relationship structure loss. Subjective and objective experiments showcase the effectiveness of our approach. Furthermore, experiments conducted across a variety of tasks, including novel view synthesis, pedestrian intention estimation, and visual grounding, demonstrate that VRT indeed provides supplementary cues for tasks related to novel views.

References

1. Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. International Journal of Computer Vision **120**, 153–168 (2016)
2. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.: Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 422–440. Springer (2020)
3. Azizian, W., Lelarge, M.: Expressive power of invariant and equivariant graph neural networks. arXiv preprint arXiv:2006.15646 (2020)
4. Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J.P., Kornbluth, M., Molinari, N., Smidt, T.E., Kozinsky, B.: E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nature communications **13**(1), 2453 (2022)
5. Bronstein, M.M., Bruna, J., Cohen, T., Veličković, P.: Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint arXiv:2104.13478 (2021)
6. Cai, B., Huang, J., Jia, R., Lv, C., Fu, H.: Neuda: Neural deformable anchor for high-fidelity implicit surface reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2023)
7. Chang, A., Savva, M., Manning, C.D.: Learning spatial knowledge for text to 3d scene generation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2028–2038 (2014)
8. Chen, S., Shi, Z., Mettes, P., Snoek, C.G.: Social fabric: Tubelet compositions for video relation detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13485–13494 (2021)
9. Chen, T., Yu, W., Chen, R., Lin, L.: Knowledge-embedded routing network for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6163–6171 (2019)
10. Cohen, T., Weiler, M., Kicanaoglu, B., Welling, M.: Gauge equivariant convolutional networks and the icosahedral cnn. In: International conference on Machine learning. pp. 1321–1330. PMLR (2019)
11. Cohen, T., Welling, M.: Group equivariant convolutional networks. In: International conference on machine learning. pp. 2990–2999. PMLR (2016)
12. Cohen, T.S., Geiger, M., Weiler, M.: Intertwiners between induced representations (with applications to the theory of equivariant neural networks). arXiv preprint arXiv:1803.10743 (2018)
13. Cohen, T.S., Geiger, M., Weiler, M.: A general theory of equivariant cnns on homogeneous spaces. Advances in neural information processing systems **32** (2019)
14. Cohen, T.S., Welling, M.: Steerable cnns. arXiv preprint arXiv:1612.08498 (2016)
15. Dai, B., Zhang, Y., Lin, D.: Detecting visual relationships with deep relational networks. In: Proceedings of the IEEE conference on computer vision and Pattern recognition. pp. 3076–3086 (2017)
16. Deng, C., Litany, O., Duan, Y., Poulenard, A., Tagliasacchi, A., Guibas, L.J.: Vector neurons: A general framework for so (3)-equivariant networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12200–12209 (2021)
17. Desai, C., Ramanan, D., Fowlkes, C.C.: Discriminative models for multi-class object layout. International journal of computer vision **95**(1), 1–12 (2011)

18. Eslami, S.A., Jimenez Rezende, D., Besse, F., Viola, F., Morcos, A.S., Garnelo, M., Ruderman, A., Rusu, A.A., Danihelka, I., Gregor, K., et al.: Neural scene representation and rendering. Science **360**(6394), 1204–1210 (2018)
19. Farhadi, A., Sadeghi, M.A.: Phrasal recognition. IEEE transactions on pattern analysis and machine intelligence **35**(12), 2854–2865 (2013)
20. Fisher, M., Savva, M., Hanrahan, P.: Characterizing structural relationships in scenes using graph kernels. In: ACM SIGGRAPH 2011 papers, pp. 1–12 (2011)
21. Fu, H., Cai, B., Gao, L., Zhang, L.X., Wang, J., Li, C., Zeng, Q., Sun, C., Jia, R., Zhao, B., et al.: 3d-front: 3d furnished rooms with layouts and semantics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10933–10942 (2021)
22. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: International conference on machine learning (ICML). pp. 1243–1252. PMLR (2017)
23. Gupta, A., Davis, L.S.: Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In: European conference on computer vision. pp. 16–29. Springer (2008)
24. Hu, D., Zhang, Z., Hou, T., Liu, T., Fu, H., Gong, M.: Multiscale representation for real-time anti-aliasing neural rendering. arXiv preprint arXiv:2304.10075 (2023)
25. Insafutdinov, E., Dosovitskiy, A.: Unsupervised learning of shape and pose with differentiable point clouds. Advances in neural information processing systems **31** (2018)
26. Jarosz, W.: Efficient Monte Carlo methods for light transport in scattering media. University of California, San Diego (2008)
27. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 406–413. IEEE (2014)
28. Ji, W., Li, Y., Wei, M., Shang, X., Xiao, J., Ren, T., Chua, T.S.: Vidvrd 2021: The third grand challenge on video relation detection. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 4779–4783 (2021)
29. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3668–3678 (2015)
30. Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3907–3916 (2018)
31. Keriven, N., Peyré, G.: Universal invariant and equivariant graph neural networks. Advances in Neural Information Processing Systems **32** (2019)
32. Kopanas, G., Philip, J., Leimkühler, T., Drettakis, G.: Point-based neural rendering with per-view optimization. In: Computer Graphics Forum. vol. 40, pp. 29–43. Wiley Online Library (2021)
33. Krishna, R., Chami, I., Bernstein, M., Fei-Fei, L.: Referring relationships. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6867–6876 (2018)
34. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017)

35. Lassner, C., Zollhofer, M.: Pulsar: Efficient sphere-based neural rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1440–1449 (2021)
36. Levoy, M., Hanrahan, P.: Light field rendering. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 31–42 (1996)
37. Li, R., Zhang, S., Wan, B., He, X.: Bipartite graph network with adaptive message passing for unbiased scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11109–11119 (2021)
38. Li, Y., Yang, X., Shang, X., Chua, T.S.: Interventional video relation detection. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 4091–4099 (2021)
39. Li, Y., Ouyang, W., Wang, X., Tang, X.: Vip-cnn: Visual phrase guided convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1347–1356 (2017)
40. Liang, K., Guo, Y., Chang, H., Chen, X.: Visual relationship detection with deep structural ranking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
41. Liang, X., Lee, L., Xing, E.P.: Deep variation-structured reinforcement learning for visual relationship and attribute detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 848–857 (2017)
42. Lin, C.H., Kong, C., Lucey, S.: Learning efficient point cloud generation for dense 3d object reconstruction. In: proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
43. Lin, X., Ding, C., Zeng, J., Tao, D.: Gps-net: Graph property sensing network for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3746–3753 (2020)
44. Liu, B., Dong, Q., Hu, Z.: Hardness sampling for self-training based transductive zero-shot learning. In: CVPR. pp. 16499–16508 (2021)
45. Liu, B., Hu, L., Hu, Z., Dong, Q.: Hardboost: Boosting zero-shot learning with hard classes. arXiv preprint arXiv:2201.05479 (2022)
46. Liu, C., Jin, Y., Xu, K., Gong, G., Mu, Y.: Beyond short-term snippet: Video relation detection with spatio-temporal global context. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10840–10849 (2020)
47. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: Learning dynamic renderable volumes from images. arXiv preprint arXiv:1906.07751 (2019)
48. Loper, M.M., Black, M.J.: Opendr: An approximate differentiable renderer. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13. pp. 154–169. Springer (2014)
49. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 852–869. Springer (2016)
50. Ma, L., Li, X., Liao, J., Zhang, Q., Wang, X., Wang, J., Sander, P.V.: Deblurnerf: Neural radiance fields from blurry images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12861–12870 (2022)
51. Mildenhall, B., Hedman, P., Martin-Brualla, R., Srinivasan, P.P., Barron, J.T.: Nerf in the dark: High dynamic range view synthesis from noisy raw images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16190–16199 (2022)

52. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
53. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3504–3515 (2020)
54. Plummer, B.A., Mallya, A., Cervantes, C.M., Hockenmaier, J., Lazebnik, S.: Phrase localization and visual relationship detection with comprehensive image-language cues. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1928–1937 (2017)
55. Qi, M., Li, W., Yang, Z., Wang, Y., Luo, J.: Attentive relational networks for mapping images to scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3957–3966 (2019)
56. Qian, X., Zhuang, Y., Li, Y., Xiao, S., Pu, S., Xiao, J.: Video relation detection with spatio-temporal graph. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 84–93 (2019)
57. Qiu, J., Wang, X., Fua, P., Tao, D.: Matching seqlets: An unsupervised approach for locality preserving sequence matching. IEEE transactions on pattern analysis and machine intelligence **43**(2), 745–752 (2019)
58. Qiu, J., Wang, X., Maybank, S.J., Tao, D.: World from blur. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8493–8504 (2019)
59. Qiu, J., Yang, Y., Wang, X., Tao, D.: Hallucinating visual instances in total absentia. In: European Conference on Computer Vision (2020)
60. Qiu, J., Yang, Y., Wang, X., Tao, D.: Scene essence. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8322–8333 (2021)
61. Rasouli, A., Kotseruba, I., Kunic, T., Tsotsos, J.K.: Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6262–6271 (2019)
62. Rematas, K., Liu, A., Srinivasan, P.P., Barron, J.T., Tagliasacchi, A., Funkhouser, T., Ferrari, V.: Urban radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12932–12942 (2022)
63. Ren, P., Dong, Y., Lin, S., Tong, X., Guo, B.: Image based relighting using neural networks. ACM Transactions on Graphics (ToG) **34**(4), 1–12 (2015)
64. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence **39**(6), 1137–1149 (2016)
65. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
66. Roveri, R., Rahmann, L., Oztireli, C., Gross, M.: A network architecture for point cloud classification via automatic depth images generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4176–4184 (2018)
67. Sainz, M., Pajarola, R.: Point-based rendering techniques. Computers & Graphics **28**(6), 869–879 (2004)

68. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
69. Shang, X., Li, Y., Xiao, J., Ji, W., Chua, T.S.: Video visual relation detection via iterative inference. In: Proceedings of the 29th ACM international conference on Multimedia. pp. 3654–3663 (2021)
70. Shang, X., Ren, T., Guo, J., Zhang, H., Chua, T.S.: Video visual relation detection. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1300–1308 (2017)
71. Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. Advances in Neural Information Processing Systems **32** (2019)
72. Su, Z., Shang, X., Chen, J., Jiang, Y.G., Qiu, Z., Chua, T.S.: Video relation detection via multiple hypothesis association. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 3127–3135 (2020)
73. Sun, X., Ren, T., Zi, Y., Wu, G.: Video visual relation detection via multi-modal feature fusion. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 2657–2661 (2019)
74. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3716–3725 (2020)
75. Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Sunkavalli, K., Martin-Brualla, R., Simon, T., Saragih, J., Nießner, M., et al.: State of the art on neural rendering. In: Computer Graphics Forum. vol. 39, pp. 701–727. Wiley Online Library (2020)
76. Tsai, Y.H.H., Divvala, S., Morency, L.P., Salakhutdinov, R., Farhadi, A.: Video relationship reasoning using gated spatio-temporal energy graph. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10424–10433 (2019)
77. Turki, H., Ramanan, D., Satyanarayanan, M.: Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12922–12931 (2022)
78. Wang, C., Wu, X., Guo, Y.C., Zhang, S.H., Tai, Y.W., Hu, S.M.: Nerf-sr: High quality neural radiance fields using supersampling. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 6445–6454 (2022)
79. Wang, J., Dong, Y., Tong, X., Lin, Z., Guo, B.: Kernel nyström method for light transport. In: ACM SIGGRAPH 2009 papers, pp. 1–10 (2009)
80. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
81. Weiler, M., Cesa, G.: General e (2)-equivariant steerable cnns. Advances in neural information processing systems **32** (2019)
82. Weiler, M., Geiger, M., Welling, M., Boomsma, W., Cohen, T.S.: 3d steerable cnns: Learning rotationally equivariant features in volumetric data. Advances in Neural Information Processing Systems **31** (2018)
83. Wiles, O., Gkioxari, G., Szeliski, R., Johnson, J.: Synsin: End-to-end view synthesis from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7467–7477 (2020)
84. Xiangli, Y., Xu, L., Pan, X., Zhao, N., Rao, A., Theobalt, C., Dai, B., Lin, D.: Citynerf: Building nerf at city scale. arXiv preprint arXiv:2112.05504 (2021)

85. Xu, X., Qiu, J., Wang, X., Wang, Z.: Relationship spatialization for depth estimation. In: European Conference on Computer Vision. pp. 615–637. Springer (2022)
86. Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph r-cnn for scene graph generation. In: Proceedings of the European conference on computer vision (ECCV). pp. 670–685 (2018)
87. Yang, Y., Qiu, J., Song, M., Tao, D., Wang, X.: Distilling knowledge from graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
88. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 17–24. IEEE (2010)
89. Yin, G., Sheng, L., Liu, B., Yu, N., Wang, X., Shao, J., Loy, C.C.: Zoom-net: Mining deep feature interactions for visual relationship recognition. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 322–338 (2018)
90. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021)
91. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4578–4587 (June 2021)
92. Yu, H., Li, R., Xie, S., Qiu, J.: Shadow-enlightened image outpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7850–7860 (2024)
93. Yu, R., Li, A., Morariu, V.I., Davis, L.S.: Visual relationship detection with internal and external linguistic knowledge distillation. In: Proceedings of the IEEE international conference on computer vision. pp. 1974–1982 (2017)
94. Zhang, H., Kyaw, Z., Chang, S.F., Chua, T.S.: Visual translation embedding network for visual relation detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5532–5540 (2017)
95. Zhang, J., Kalantidis, Y., Rohrbach, M., Paluri, M., Elgammal, A., Elhoseiny, M.: Large-scale visual relationship understanding. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 9185–9194 (2019)
96. Zhang, J., Huang, J., Cai, B., Fu, H., Gong, M., Wang, C., Wang, J., Luo, H., Jia, R., Zhao, B., et al.: Digging into radiance grid for real-time view synthesis with detail preservation. In: European Conference on Computer Vision. pp. 724–740. Springer (2022)
97. Zheng, S., Chen, S., Jin, Q.: Vrdformer: End-to-end video visual relation detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18836–18846 (2022)
98. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)
99. Zhuang, B., Liu, L., Shen, C., Reid, I.: Towards context-aware interaction recognition for visual relationship detection. In: Proceedings of the IEEE international conference on computer vision. pp. 589–598 (2017)
100. Zwicker, M., Pfister, H., Van Baar, J., Gross, M.: Surface splatting. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 371–378 (2001)
101. Zwicker, M., Pfister, H., Van Baar, J., Gross, M.: Ewa splatting. IEEE Transactions on Visualization and Computer Graphics **8**(3), 223–238 (2002)