# Relationship Spatialization for Depth Estimation

Xiaoyu Xu[1], Jiayan Qiu[1], Xinchao Wang[2], and Zhou Wang[1]

[1] University of Waterloo, Canada
[2] National University of Singapore, Singapore
x423xu@uwaterloo.ca, jiayan.qiu.1991@outlook.com, xinchao@nus.edu.sg,
zhou.wang@uwaterloo.ca

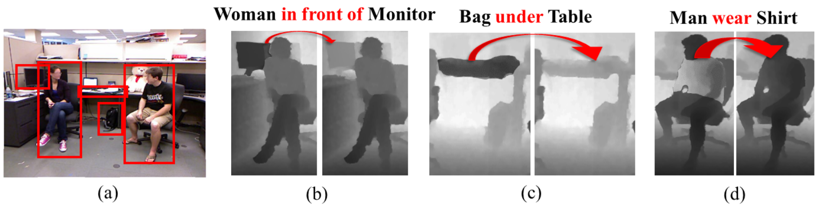

Fig. 1: Given a scene image (a), relationships contribute to depth estimation, as shown left to right in (b-d). Compared with the explicit spatial information from *in front of* (a), the implicit spatial information from *under* (b) also benefits the depth estimation. More interestingly, *wear* (c), which is intuitively non-spatial, in reality contributes significantly to depth estimation.

**Abstract.** Considering the role played by the inter-object relationships in monocular depth estimation (MDE), it is easy to tell that relationships, such as *in front of* and *behind*, provide explicit spatial priors. However, it is hard to answer the questions that which relationships contain useful spatial priors for depth estimation, and how much do their spatial priors contribute to the depth estimation? In this paper, we term the task of answering these two questions as 'Relationship Spatialization' for Depth Estimation. To this end, we strive to spatialize the relationships by devising a novel learning-based framework. Specifically, given a scene image, its image representations and relationship representations are first extracted. Then, the relationship representations are modified by spatially aligned into the visual space and redundancy elimination. Finally, the modified relationship representations are adaptively weighted to concatenate with the image ones for depth estimation, thus accomplishing the relationship spatialization. Experiments on KITTI, NYU v2, and ICL-NUIM datasets show the effectiveness of the relationship spatialization on MDE. Moreover, adopting our relationship spatialization framework to the current state-of-the-art MDE models leads to marginal improvement on most evaluation metrics.

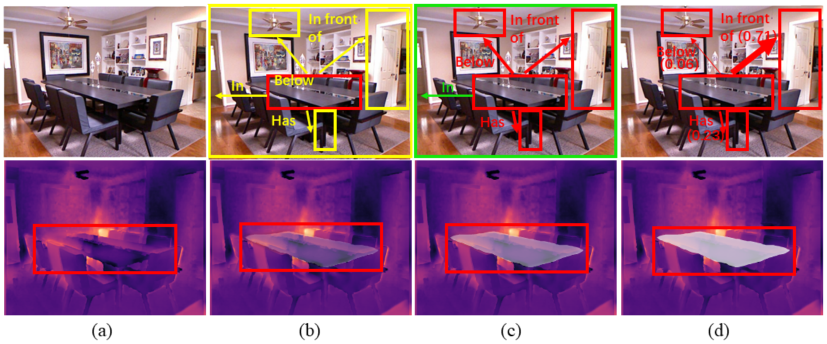

Fig. 2: Depth error decreases during our relationship spatialization process in the red boxed Table area (a-d). Only the visual understanding is utilized in (a); all Table-related in-image relationships are added in (b); only the effective relationships are identified and then equally added in (c); the identified relationships are contribution-accordingly weighted and then added(d).

## 1 Introduction

Having a look at Fig. 1, it can be found that the depth estimation benefits from cooperating with the relationships in the scene. For example, the woman is closer to us than the screen because it is *in front of* the screen, which provides explicit spatial prior. The table is farther than the bag, although the spatial prior of *under* is implicit, the bag is with smaller depth in the imaging coordinate because it is closer to the image center compared with the table. More interestingly, although *wear* seems with no spatial priors intuitively, it certainly contributes to the depth estimation, since it constrains that the depth of the shirt should be aligned with the depth of the man. Hence, it can be seen that the inter-object relationships provide a huge amount of spatial priors for depth estimation.

Despite the relationships come with lots of spatial priors, before we claim their effects on MDE, there are two questions that require explorations. Firstly, which kinds of relationships are effective for MDE? Some of the relationships always come with spatial priors, such as *behind*, while others, like *has*, may not. Thus, it is important to learn to find effective relationships for each object. Secondly, how much do the effective ones contribute to MDE? Even human observers can easily find out which relationships provide the cues for the object's depth, it is indeed hard to tell their relative importance. Therefore, in order to reveal the embedded spatial priors in the relationships, it is important to quantify their contribution to MDE. In this paper, we term the task of answering these two questions as *Relationship Spatialization for Depth Estimation*.

Towards solving the proposed Relationship Spatialization task, we propose a novel framework that combines both visual understanding and relationship understanding for depth estimation. Different from the simple combination of information, the relationship representations are automatically and adaptively weighted, thus are learned to be effectiveness-identified and contribution-quantified. Thus, the relationships are spatialized for depth estimation.

Specifically, given a scene image, we firstly feed it into an image representation extractor module, which can be any deep learning MDE model, to obtain the image representations. Meanwhile, the objects in the image are detected to construct the corresponding scene graph, where nodes denote the objects and edges denote the inter-object relationships. Then, the scene graph is processed by the relationship recognition module to predict the inter-object relationships and extract their corresponding representation. Secondly, the relationship representations from the scene graph space are modified by spatially aligning into the visual space and redundancy elimination. Finally, the modified relationship representations are fed into an attention layer for representation weighting, which is the respective contribution quantification. Then, the concatenation of the weighted relationship representations and the image ones is fed into the depth predictor for depth estimation. Through this process, the relationships are spatialized for enhancing the depth estimation, as shown in Fig. 2.

Our contribution is therefore a novel framework that, for the first time, tries to spatialize the relationships for depth estimation, during which the spatial priors from relationships are mined automatically, identified effectively, and quantified adaptively. Moreover, our proposed framework can be implemented on any deep learning-based MDE method for performance boosting. Experiments on all datasets show encouraging and promising results.

## 2 Related work

In this section, we briefly review the prior works related to ours, including depth estimation, relationship recognition and graph neural network.

**Depth Estimation.** Early works of depth estimation [87, 22, 21] focus on utilizing the geometry-based information, such as the points correspondence among different views. With the development of deep learning techniques, the CNN based depth methods [18, 17, 48, 7, 24, 5] show high-performance on depth estimation tasks. Some of them try to improve the performance by avoiding using the ground-truth information from the real world [85, 118, 2, 13, 117, 29, 32, 119, 93, 82, 25, 23, 26, 63], which reduce the demand for expensive annotations. Other works aim to increase the methods' generalizability and work in the wild [121, 11, 98, 12, 99, 58, 71] thus saving the heavy training cost. The most related works to ours are the multi-task ones[38, 60, 41, 47, 17, 95, 53, 85, 42, 76, 100, 115, 116, 10, 68, 61], where information from other tasks are used in depth estimation. However, none of them try to utilize the spatial priors in inter-object relationships.

**Relationship Recognition.** The handful of visual relationships are categorized into three types: the inter-object ones [16], the property ones [34] and the activity ones [106]. In this work, we focus on the inter-object relationships, where the early works focus on the visual phrase recognition [57, 19, 55]. Then, thanks to the boosting performance brought by deep learning, the DNN based methods show promise performance due to their high representation capability

[14,67,111,113,120,45,56,75,62,64]. The most related works to ours are the ones that try to detect the relationship in scene graphs as one of the attributes [57,40,20,8,59]. However, none of these works explore the spatial priors in the relationships and their enhancement on depth estimation.

**Graph Neural Network.** Earlier works on graph-related tasks usually require pre-defined node and edge features [77,96,97,70,69,49,78], or aggregating the node and edge features in an iterative manner, which is computation expensive and time consuming [27,86,33,89]. Recently, graph neural networks have been proposed to learn the features. Specifically, the spectral-based networks implement the spectral theory for graph convolution design [54,52,44,15,37, 6], and spatial-based ones utilize the mutual dependency by designing the information aggregation manner [92,39,9,4,28,72,35,31,73,3,1,51,66,104,103, 102,114,91,80,79,112,84,65,108,109,94,105]. More recently, the hierarchical GCN [107] with high representation capability is proposed and shows promising performance.

## 3 Method

In this section, we explain the proposed framework in detail. As depicted in Fig. 3, our framework comprises three stages. In Stage 1, the image representations of the input image are extracted from a deep learning-based MDE model. Meanwhile, a detection module is utilized on the input image to detect the objects in the scene, which constructs the scene graph. Afterward, the relationship representations are extracted from the scene graph by implementing a relationship recognition module. In Stage 2, we first spatially align the relationship representations from the graph space with the image representations from the visual space. Then, the aligned relationship representations are modified by eliminating their redundancies compared with the image representations. In Stage 3, the concatenation of the modified relationship representation and the image representation is fed into the depth predictor the depth estimation, during which the attention is implemented on the modified relationship representations. The attention scores are automatically learned and used for identifying the effective relationships and quantifying their respective contributions on MDE.

### 3.1 Stage 1: Representation Extraction

**Image Representation Extraction.** The image representation extractor is implemented to obtain understanding from appearance cues, such as edges and shapes while eliminating the interference from surface textures. As shown in Fig.4, a U-Net encode-decoder network is adopted as our image representation extractor, where different resolution representations are combined by BiFPN[90] to preserve both the high-level understanding and visual details. Then, the image representation is:

$$R_I = f_{image}(I), \tag{1}$$

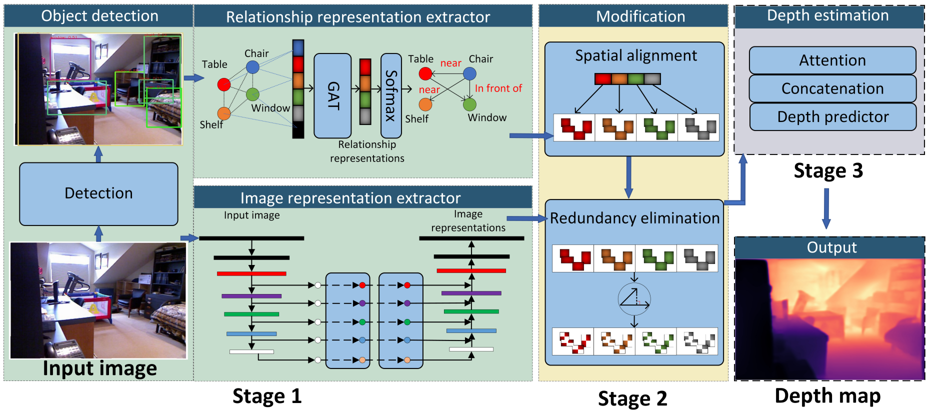

Fig. 3: The framework of the proposed model. An input image is firstly sent to an image feature extractor to generate image representations. The input image is also fed to relationship recognition module to generate relationship representations. then, the relationship ones are aligned to the image ones and being eliminated with redundancy. Finally, the effective relationship representations are quantified and concatenated with image features to estimate depth.

where $I$ denotes the input image and $f_{image}$ denotes the image representation extractor.

**Relationship Representation Extraction.** We first adopt an offline detector model [83] to detect the objects in the scene image so as to derive object-level understanding and their spatial position:

$$z_i, M_i = Detector(I), i = 1, .., N, \tag{2}$$

where $N$ denotes the number of the detected objects and $z_i$ denotes the object-level understanding of the $i$-th detected object. $M_i$ denotes the object mask, 1 for the object area. Then, the graph-rcnn [101] is adopted as the relationship recognition module. Specifically, the object-level understanding $(z_1, z_2, \cdots, z_N)$ of the $N$ detected objects are treated as nodes features of the scene graph $G$. The $N \times N$ edges in the $G$ correspond to $N \times N$ $< subject, object >$ relationships. Then, the node features are aggregated as:

$$z_i^{(l+1)} = \sigma\left(z_i^{(l)} + \sum_{i \in N(i)} a_{ij} W z_j^{(l)}\right), \tag{3}$$

$$a_{ij} = softmax(\sigma(W_a[z_i^{(l)}, z_j^{(l)}])), \tag{4}$$

where $z_i^{(l)}$ denotes the $i$-th node at $l$-th iteration, $\sigma$ denotes activation function, $N(i)$ denotes neighboring nodes of the $i$-th node, $W$ and $W_a$ are learnable parameters, and $a_{ij}$ denotes attention between the $i$-th and $j$-th nodes. Afterward, the relationship representations are computed as:

$$R_r^{i,j} = embedding([z_i^{(l)}.z_j^{(l)}]). \tag{5}$$

Due to the deep learning based depth estimation methods are only able to deal with fixed sized input, we extract the fixed number of $N_r$ relationship repre-

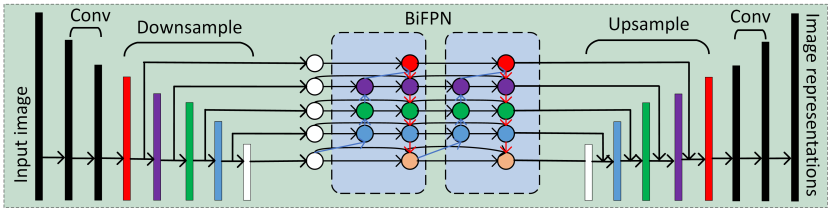

Fig. 4: Image representation extractor

sentations from the relationship recognition module. Specifically, we feed the relationship representations $R_r^{i,j}$ into a relationship predictor and preserve the ones with top-$N_r$ prediction scores.

### 3.2 Stage 2: Representation Modification

After obtaining the image representations $R_I$ and relationship representations $R_r^n$, where $n \in [1, N_r]$, we modify the relationship ones according to the image ones by first spatial alignment and followed by the redundancy elimination.

**Spatial Alignment.** Since the image representations, $R_I$ are in the visual space and the relationship representations $R_r^n$ are in the graph space, spatial alignment is implemented to translate the relationship ones into the visual space for depth estimation. Specifically, given the relationship representation $R_r^n \in \mathbb{R}^d$, where $d$ denotes the representation dimension of $R_r^n$, and its corresponding object masks $M_i$ and $M_j$, we first project the relationship representation upon a zero-initialized feature set:

$$F^n(x, y, :) = R_r^n(M_i(x, y)||M_j(x, y)), x = [1, ..., I_w], y = [1, ..., I_h], \tag{6}$$

where $F^n \in \mathbb{R}^{I_w \times I_h \times d}$, $I_w$ and $I_h$ denote the width and height of the input image $I$, and $||$ denotes the logic operation *or*. Then, a single output channel convolutional layer is implemented on the feature set $F$ to obtain the visually aligned relationship representation:

$$VA\_R^n = Conv(F^n), n = 1, ..., N_r, \tag{7}$$

where $VA\_R^n \in \mathbb{R}^{I_w \times I_h}$ denotes the visually aligned single channel feature map of each relationship representation.

**Redundancy Elimination.** After obtaining the visually aligned feature maps of relationship representations, we eliminate their redundancy information according to the image representations, since the overlap information between them may lead to overlook the distinguished spatial information contained in the relationships. In detail, the overlaps between the relationship representation and the image representation are eliminated by:

$$M_R^c = G(M_R^{c-1}, R_I^{c-1}), c = 1, ..., N_I, \tag{8}$$

$$G(M_R^{c-1}, R_I^{c-1}) = M_R^{c-1} - \frac{< M_R^{c-1}, R_I^{c-1} >}{< R_I^{c-1}, R_I^{c-1} >} R_I^{c-1}, \tag{9}$$

$$M_R^0 = VA\_R, \tag{10}$$

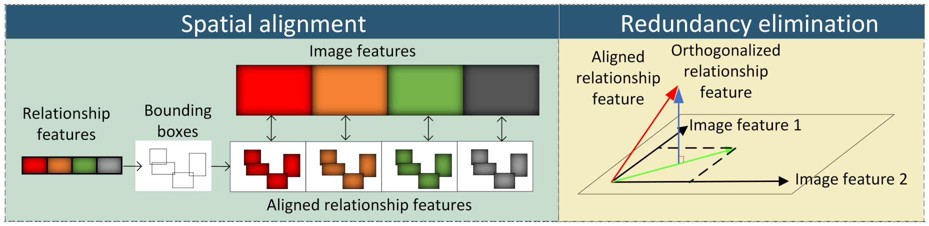

Fig. 5: Representation modification is composed of spatial alignment and redundancy elimination. Firstly, relationship representations are spatially aligned towards image representations with bounding boxes. Secondly, the aligned relationship are orthogonalized against image representations to eliminate information redundancy.

where $M_R$ denotes the modified relationship representation, $N_I$ denotes the number of feature maps in the image representation $R_I$ and $< \cdot, \cdot >$ denotes inner product. At the end of this stage, we obtain the modified relationship representation $M_R$.

### 3.3 Stage 3: Depth Estimation and Contribution Quantification

Once the the modified relationship representations $M_R$ are obtained, we feed them into an attention layer for learning to identify their effectiveness and quantify their respective contributions:

$$\alpha^n = Attn(M_R^n), n = 1, ..., N_r, \tag{11}$$

where the attention scores $\alpha^n$ are used as the effectiveness identification and contribution quantification of the relationships. Then, we concatenate the image representations and the weighted relationship representations, feed it in the depth estimator for depth estimation:

$$y = Predictor(Concat(R_I, \alpha^n M_R^n)), n = 1, ..., N_r. \tag{12}$$

Then we impose the scale invariant loss [5], who focus on minimizing the ratio of the depth error, rather than the absolutely scaled depth:

$$Ł_{depth} = var([\log(y) - \log(\tilde{y})]) + \lambda[\frac{1}{N}\sum_{i=1}^{N}[\log(y) - \log(\tilde{y})]]^2, \tag{13}$$

where $var()$ denotes variance computation, $\tilde{y}$ denotes the ground-truth depth and $\lambda$ denotes the balancing weight.

## 4 Experiments

In this section, we provide our experimental setups and show the results. Specifically, for the experiments, we first evaluate the performance of our framework on three MDE datasets. Then analysis the effectiveness of relationship spatialization, effective spatialized relationship identification and the respective contribution quantification are explained in detail. Finally, ablation study is conducted for measuring the necessary of the redundancy elimination operation.

Our goal is, again, spatializing the inter-object relationships, then identifying the effective ones and quantifying their respective contributions on MDE. Since we are not aware of any existing work that performs same task, we aim to show the promise of the proposed framework, rather than trying to beat any state-of-the-art monocular depth estimation and relationship recognition models. More complicated networks, as long as they are end-to-end trainable, can be adopted to our framework to achieve potentially better results. More results are shown in supplementary materials.

### 4.1 Datasets and Implementation details

We adopt three datasets for depth estimation, KITTI [30], NYU Depth v2 [88], and ICL-NUIM [36] to validate the proposed framework. Since there is no dataset that provides both the ground-truth annotated depth and inter-object relationship, we evaluate our framework by the MDE performance. As for the relationship recognition, we pretrain a model on Visual Genome dataset[46], and adopt it directly in our framework. User studies are designed for evaluating the performance of the adopted relationship recognition module on the three depth estimation datasets.

**KITTI[30].** It is a dataset of outdoor scene images with the corresponding 3D laser scans captured by moving vehicles. All the scene categories are used as our training/test sets. Following the split manner in [18], we use 26k images for training and 697 images for the test, with the image resolution of 1241 × 376. For a fair comparison, all compared MDE methods are retrained and evaluated on the adopted training/test sets in our experiments.

**NYU Depth v2 [88].** It is a dataset of indoor scene images with the corresponding depth maps captured by Microsoft Kinect. Following the previous work, we use 50k images from 249 scenes for training and the 654-images set for the test, with the image resolution of 640 × 480. Different from the existing MDE methods like [5] that training only on cropped parts of the image, which may lead to the loss of relationship information, we perform the training on the entire image.

**ICL-NUIM [36].** It is a smaller size dataset of an indoor scene, which comprises of 1500 images with their corresponding depth maps. Due to its limited amount of images, we adopt this dataset for verifying the generalizability of the model trained on NYU Depth v2.

**Implementation.** In the experiment, our framework is implemented with *Pytorch* [74] and with 4 NVIDIA P100 Pascal GPUs. We optimize with Adam [43], setting the learning rate $l = 5e-4$, $\beta_1 = 0.9$, and $\beta_2 = 0.99$. Our framework is trained with batch size 16 for 25 epochs.

### 4.2 Performance Evaluation

In this part, we evaluate the performance of our proposed relationship spatialization framework on MDE. In order to show the effect of the learned relationship spatialization, we conduct the evaluations with our baseline framework and

| Metrics | | Adabins | Adabins+ | Midas | Midas+ | Bts | Bts+ | Baseline | Baseline + |
|---|---|---|---|---|---|---|---|---|---|
| ↑ | $\delta_1$ | 0.9292 | **0.9483** | 0.9483 | **0.9486** | 0.8832 | **0.9161** | 0.9113 | **0.9260** |
| | $\delta_2$ | 0.9883 | **0.9911** | 0.9912 | **0.9927** | 0.9746 | **0.9837** | 0.9852 | **0.9884** |
| | $\delta_3$ | 0.9973 | **0.9981** | 0.9980 | **0.9984** | 0.9937 | **0.9960** | 0.9972 | **0.9977** |
| ↓ | *abs_rel* | 0.0775 | **0.0758** | 0.0655 | **0.0652** | 0.1195 | **0.0831** | 0.0917 | **0.0789** |
| | *rmse* | 3.4291 | **2.9441** | **2.8099** | 2.8747 | 4.4126 | **3.4930** | 3.6264 | **3.2560** |
| | *log_10* | 0.0348 | **0.0339** | 0.0291 | **0.0289** | 0.0486 | **0.0368** | 0.0399 | **0.0342** |
| | *rmse_log* | 0.1196 | **0.1104** | 0.1026 | **0.1025** | 0.1567 | **0.1297** | 0.1327 | **0.1191** |
| | *silog* | 10.6605 | **9.5903** | 9.3816 | **9.3112** | **11.8858** | 13.5749 | 12.3885 | **11.0609** |
| | *sq_rel* | 0.3406 | **0.2875** | 0.2660 | **0.2603** | 0.4026 | **0.4026** | 0.4042 | **0.3554** |

Table 1: Results of monocular depth estimation on KITTI dataset. ↑: the higher is better. ↓: the lower is better.

| Metrics | Adabins | Adabins+ | Midas | Midas+ | Bts | Bts+ | Baseline | Baseline+ |
|---|---|---|---|---|---|---|---|---|
| $\delta_1$ | 0.7971 | **0.8603** | 0.8173 | **0.8519** | 0.6936 | **0.7715** | 0.8293 | **0.8889** |
| $\delta_2$ | 0.9486 | **0.9704** | 0.9519 | **0.9655** | 0.9214 | **0.9493** | 0.9522 | **0.9597** |
| $\delta_3$ | 0.9840 | **0.9928** | 0.9858 | **0.9901** | 0.9803 | **0.9852** | 0.9791 | **0.9894** |
| *abs_rel* | 0.1537 | **0.1203** | 0.1437 | **0.1282** | 0.1772 | **0.1557** | 0.1488 | **0.1432** |
| *rmse* | 0.5318 | **0.4469** | 0.5016 | **0.4717** | 0.6652 | **0.5847** | 0.7632 | **0.5342** |
| *log_10* | 0.0629 | **0.0515** | 0.0596 | **0.0544** | 0.0806 | **0.0703** | **0.0596** | 0.0646 |
| *rmse_log* | 0.1900 | **0.1588** | 0.1822 | **0.1779** | 0.2327 | **0.2245** | 0.2015 | **0.1924** |
| *silog* | 15.9615 | **13.9534** | 15.4977 | **15.4224** | **18.5223** | 21.9191 | 17.9495 | **15.8711** |
| *sq_rel* | 0.1234 | **0.0803** | 0.1116 | **0.0896** | 0.1542 | **0.1233** | 0.1979 | **0.1032** |

Table 2: Results of monocular depth estimation on NYU Depth v2 dataset.

frameworks that implement the commonly used MDE networks as the image representation extractor. In the experiment, the state-of-the-art MDE methods, Adabins [5], Midas [81] and Bts [50], are adopted. We use Baseline, Adabins, Midas, and Bts to denote the MDE-only frameworks; and use Baseline+, Adabins+, Midas+ and Bts+ to denote the frameworks with relationship spatialization. For a fair comparison, all frameworks in the experiment are trained from the scratch.

**KITTI.** It can be seen from Table 1 that, for all chosen frameworks on KITTI dataset, relationship spatialization brings performance gain on almost every MDE evaluation metric, which shows that our proposed framework is able to learn to find the effective relationship spatialization for the outdoor scenes. Moreover, it is worth noting that the framework Bts+ performs obviously poor than Bts on the *silog* metric. Since the *silog* reflects the variation of scale error (calculated by dividing predicted depth with ground truth depth), the *rmse_log* reflects the mean value of the scale error, the BTS+ model focuses more on reducing mean error in relationship related areas and neglects areas without relationship enhancement, thus bringing in variation increment. So the BTS+ gets the greatest progress in *rmse_log* and does not outperform in *silog*.

**NYU Depth v2.** The results on NYU Depth v2 dataset are shown in Table 2, where the frameworks with the relationship spatialization outperform the original ones on almost all of the evaluation metrics. It shows that our proposed framework is able to learn to find the effective relationship spatialization of the indoor scenes. It is worth noting that, again, the Bts outperforms the Bts+ in the *silog* metric due to the same reason on KITTI.

| Metrics | Adabins | Adabins+ | Midas | Midas+ | Bts | Bts+ | Baseline | Baseline + |
|---|---|---|---|---|---|---|---|---|
| $\delta_1$ | 0.9160 | **0.9521** | **0.8432** | 0.7903 | 0.6735 | **0.7758** | 0.7344 | **0.8446** |
| $\delta_2$ | 0.9993 | **1.0000** | 0.9446 | **0.9566** | 0.9112 | **0.9539** | 0.9438 | **0.9744** |
| $\delta_3$ | 1.0000 | **1.0000** | **1.0000** | 0.9890 | 0.9783 | **0.9882** | 0.9902 | **0.9943** |
| *abs_rel* | 0.1048 | **0.0758** | 0.1878 | **0.1521** | 0.2075 | **0.1634** | 0.1732 | **0.1297** |
| *rmse* | 0.3283 | **0.2584** | 0.5462 | **0.3492** | 0.4727 | **0.3726** | 0.4158 | **0.3141** |
| *log_*10 | 0.0438 | **0.0319** | 0.0761 | **0.0637** | 0.0843 | **0.0694** | 0.0709 | **0.0545** |
| *rmse_log* | 0.1276 | **0.0959** | 0.2059 | **0.1936** | 0.2417 | **0.2373** | 0.2076 | **0.1681** |
| *silog* | 12.6962 | **8.5677** | 20.8457 | **19.2291** | 24.0221 | **23.6079** | 20.5816 | **16.7161** |
| *sq_rel* | 0.0438 | **0.0261** | 0.1273 | **0.0770** | 0.1456 | **0.0827** | 0.1026 | **0.0578** |

Table 3: Results of frameworks trained on NYU Depth v2 dataset and evaluated on ICL_NUIM dataset.

| Models | | $\delta_1$ | $\delta_2$ | $\delta_3$ | *abs_rel* | *rmse* | *log_*10 | *rmse_log* | *silog* | *sq_rel* |
|---|---|---|---|---|---|---|---|---|---|---|
| KITTI | B | 0.9292 | 0.9883 | 0.9973 | 0.0775 | 3.4291 | 0.0348 | 0.1196 | 10.6605 | 0.3406 |
| | +R | **0.9446** | **0.9910** | **0.9981** | **0.0738** | **2.9988** | **0.0331** | **0.1098** | **9.5838** | **0.2892** |
| NYU | B | 0.8237 | 0.9597 | 0.9895 | 0.1390 | 0.5006 | 0.0581 | 0.1770 | 15.0716 | 0.1042 |
| | +R | **0.8608** | **0.9715** | **0.9932** | **0.1295** | **0.4487** | **0.0519** | **0.1576** | **13.2523** | **0.0796** |

Table 4: The model added with relationship representations (+R) outperforms baseline model only using image representations (B) both on KITTI and NYU datasets.

**ICL_NUIM.** Due to its limited amount of training data, ICL_NUIM dataset is utilized for measuring the generalizability of our proposed framework. Specifically, we pre-train all frameworks on NYU Depth v2 dataset and evaluate them on ICL_NUIM dataset. As can be seen from Table 3, the frameworks with relationship spatialization outperform the original ones on all evaluation metrics, which shows the promising generalizability of our framework.

### 4.3 Effect of Relationship Representation

In this part, we evaluate the effect of relationship representations on MDE by performance comparison, correlation evaluation, related area evaluation, and framework derivation visualization.

**Performance Comparison.** To illustrate the effect of relationship representation for depth estimation, here we compare the performance of framework with only image representation and framework with both the image and relationship one. Note that, the attention layer in the predictor of our proposed framework is not implemented in this part, which means that all the relationship representations are equally treated. As shown in Table 4, the framework with relationship representations outperforms the original one, which shows the relationship representations benefit the depth estimation.

**Correlation Evaluation.** In this experiment, we evaluate the correlation between the MDE performance and the relationship recognition accuracy. Specifically, we adopt relationship recognition modules with different pre-trained accuracy into our framework and then observe the difference in the MDE performance. As can be seen from Table 5, we MDE performance increase with the pre-trained relationship recognition accuracy, which shows that correctly predicted relationships benefit the depth estimation.

| Relationship | | | Depth estimation | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Recall | User_Acc | | Adabins+ | | Midas+ | | Bts+ | | Baseline+ | |
| | KITTI | NYU | KITTI | NYU | KITTI | NYU | KITTI | NYU | KITTI | NYU |
| 10.5 | 79.62 | 78.58 | 2.9988 | 0.5318 | 3.0122 | 0.5134 | 7.8685 | 0.7915 | 4.5427 | 0.7463 |
| 12.5 | 81.27 | 83.01 | 2.9441 | 0.4654 | 2.9671 | 0.5016 | 7.5241 | 0.6246 | 4.1325 | 0.6824 |
| 16.7 | 84.53 | 83.27 | 2.8803 | 0.4487 | 2.9035 | 0.4925 | 7.1373 | 0.5874 | 3.9776 | 0.6596 |
| 18.3 | 88.17 | 86.31 | 2.8742 | 0.4456 | 2.8747 | 0.4791 | 6.5481 | 0.5831 | 3.8212 | 0.5342 |
| 18.8 | 89.33 | 87.25 | **2.8652** | **0.4392** | **2.8099** | **0.4717** | **6.0439** | **0.5058** | **3.2560** | **0.5209** |

Table 5: The MDE performance increases with the relationship recognition accuracy. *Recall* is used as the evaluation metric of the pre-trained relationship recognition module, *User_Acc* is used as the metric for user-oriented accuracy and *rmse* is used as the metric for MDE.

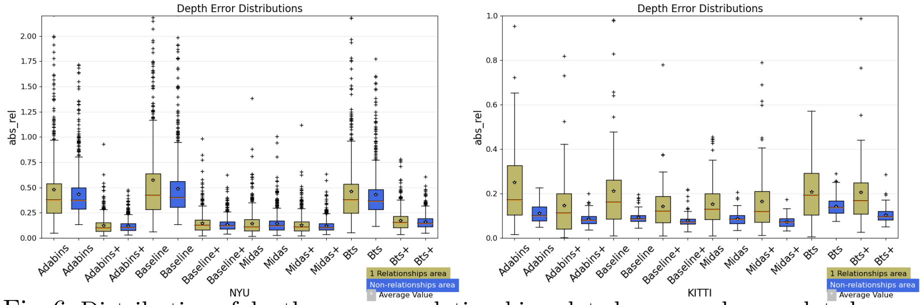

Fig. 6: Distribution of depth errors on relationship-related areas and non-related areas. The yellow boxes denote the *abs_rel* on the relationship-related areas, the blue boxes denote the *abs_rel* on the non-related areas, the in-box red line denotes median value, and the star symbol denotes the mean value.

Moreover, since there are no ground-truth annotated relationships in the MDE datasets, a subjective user study is designed for evaluating the accuracy of the pre-trained relationship recognition module on the MDE datasets. Specifically, we involve 107 users and send each user 200 randomly selected relationships, which are used in our framework. Then, we ask the user whether each relationship is correctly recognized, according to which calculate the user-oriented relationship recognition accuracy. The user-oriented accuracy is close to the ground-truth one while covering only part of the used relationships. The user-oriented accuracy is shown in the second column in Table 5, it increases with the pre-trained accuracy.

**Related Area Evaluation.** Since the representations of the spatialized relationship are spatially aligned into the visual space, the majority of its spatial priors should have effects on the relationship-related areas. Although the performance on the entire image shows improvement, it is hard to fully reveal the effect of relationship spatialization for MDE. In this experiment, we compare the MDE performance on relationship-related areas and the non-related ones, the *abs_rel* metric is chosen for clear visualization. As can be seen from the box plots in Figure. 6, the introduction of the relationship, firstly, improves both the MDE performance on the relationship-related areas and non-related areas. More importantly, the overall depth error reduction on the relationship-related areas

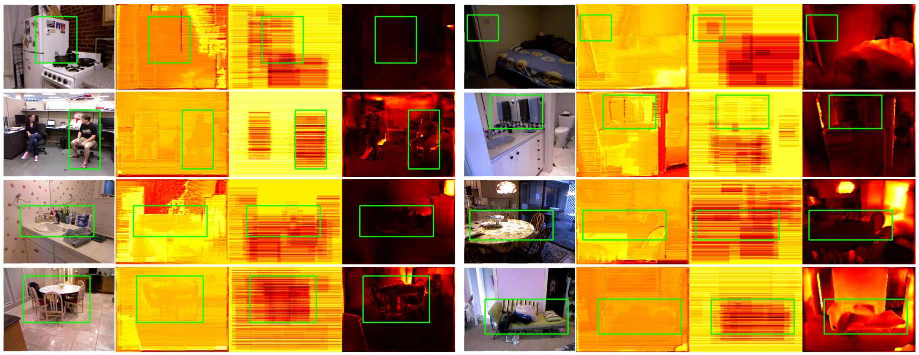

Fig. 7: Framework derivation visualization. The first column of each group shows the input scene image, the second column shows the framework derivation map, the third column shows the relationship representation map, the fourth column shows the depth error map, and the bounding box denotes the relationship area.

highly outperforms that on the non-related areas, which shows the large benefits from relationship spatialization on MDE.

**Framework Derivation Visualization.** In this experiment, we calculate and show the framework derivation map in Figure. 7. It can be seen that, after introducing the relationship representations, the derivation values are larger on the relationship areas than that on other areas, which means the relationship areas contribute more to depth estimation, thus more important. It is worth noting that, excluding the relationship areas, the sudden texture changing areas are with the highest derivation scores, which shows the fact that image representation-based methods abstract the spatial information from edge areas. Then, the introduction of our relationship spatialization helps to extract the spatial information from object surface areas, complement with the image representation-based MDE methods, thus improving the MDE performance.

### 4.4 Effective Spatialized-Relationship Identification

In this part, we evaluate the identified effective relationships from our framework by performance comparison and user study.

**Performance Evaluation.** In order to evaluate whether the identified effective relationships are meaningful for MDE, here we compare the performance of the framework with all relationship representations and the one with only the identified effective relationship representations. Specifically, a binary feature channel mask is learned from the relationship representations and used for effective relationship representations identification. Then, all identified effective relationship representations are equally treated to concatenate with the image ones for depth estimation. As shown in Table 6, the framework with only effective relationship representations outperforms the one with all relationship representations, which shows that the effective relationships for MDE are preserved and the noisy information from the non-effective relationships is eliminated.

**User Study.** Due to the lack of ground-truth annotations of the effective relationships in MDE, a user study is designed and performed to evaluate the

| Models | | $\delta_1$ | $\delta_2$ | $\delta_3$ | *abs_rel* | *rmse* | *log_10* | *rmse_log* | *silog* | *sq_rel* |
|---|---|---|---|---|---|---|---|---|---|---|
| KITTI | +R | 0.9446 | 0.9910 | 0.9981 | 0.0738 | 2.9988 | 0.0331 | 0.1098 | 9.5838 | 0.2892 |
| | +ER | **0.9512** | **0.9913** | **0.9981** | **0.0725** | **2.8574** | **0.0215** | **0.0905** | **9.4273** | **0.2138** |
| NYU | +R | 0.8608 | 0.9715 | 0.9932 | 0.1295 | 0.4487 | 0.0519 | 0.1576 | 13.2523 | 0.0796 |
| | +ER | **0.8752** | **0.9833** | **0.9932** | **0.1107** | **0.3583** | **0.0462** | **0.1113** | **12.4264** | **0.0686** |

Table 6: The model added with effective relationship representations (+ER) outperforms the model with equally treated relationship representations (+R) both on KITTI and NYU datasets.

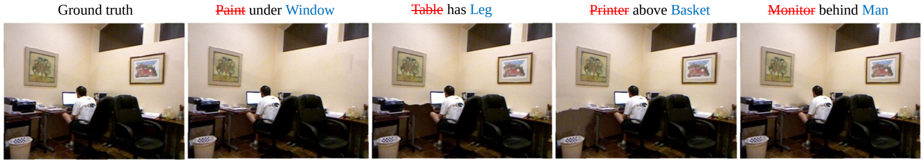

Fig. 8: In each image, one target object of the identified effective relationships is erased. Then, users' judgments on whether the erased object is effective in depth estimation are recorded.

identified effective relationships from our framework. Specifically, we involve 107 users and send each user 100 randomly selected relationship sets shown in Fig.8. Each relationship set includes the input image and 4 images that each randomly gets rid of one of our identified effective relationships, where one of the two related objects is erased with an inpainting model [110]. Then, we ask the user if the erased relationship affects the measurement of the depth of the remaining object, then evaluate if we correctly identify the effective relationships. Our identified effective relationships finally achieve 89.33% accuracy, which is highly aligned with human recognition.

### 4.5 Respective Contributions Quantification

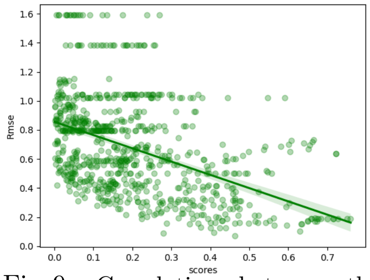

Fig. 9: Correlation between the depth error and its quantified respective contribution on relationship areas.

In this part, we evaluate the quantified respective relationship contributions from our framework by performance comparison, correlation evaluation, and user study.

**Performance Evaluation.** To evaluate whether the quantified respective contributions from our framework are accurate, here we compare the performance of the framework that equally treats all relationship representations and the one that weights relationship representations according to their quantified respective contributions, which is our proposed framework. As shown in Table 7, the framework with quantified respective contributions outperforms the one with equal contributions, which shows that our respective contribution quantification spatialized the relationship adaptively.

**Correlation Evaluation.** In this experiment, we evaluate the correlation between the depth error and our quantified respective contribution on the relationship-

| Models | | $\delta_1$ | $\delta_2$ | $\delta_3$ | *abs_rel* | *rmse* | *log_10* | *rmse_log* | *silog* | *sq_rel* |
|---|---|---|---|---|---|---|---|---|---|---|
| KITTI | +ER | 0.9512 | 0.9913 | 0.9981 | 0.0725 | 2.8574 | 0.0215 | 0.0905 | 9.4273 | 0.2138 |
| | +QR | **0.9549** | **0.9914** | **0.9981** | **0.0690** | **2.8541** | **0.0208** | **0.0856** | **9.3253** | **0.1838** |
| NYU | +ER | 0.8752 | 0.9833 | 0.9932 | 0.1107 | 0.3583 | 0.0462 | 0.1113 | 12.4264 | 0.0686 |
| | +QR | **0.8820** | **0.9833** | **0.9932** | **0.1071** | **0.3552** | **0.0413** | **0.1004** | **12.1541** | **0.0666** |

Table 7: The model added with quantified relationship representations (+QR) outperforms the model only using effective relationship representations (+ER) both on KITTI and NYU datasets.

related areas. As can be seen from Fig. 9, the relationship area depth error decrease when its quantified respective contribution increase, which shows that our respective contribution quantification accurately rank and spatialized the relationship for depth estimation.

### 4.6 Ablation Study

In this experiment, we verify the effectiveness of our redundancy elimination operation by comparing the framework with and without it. As can be seen from Table 8, the redundancy elimination operation leads to a marginal performance improvement on both the KITTI and NYU Depth v2 datasets, which is because it reduces the affection of redundancy information on the respective contribution quantification, thus increasing the focus on distinguished information from relationship spatialization.

| Models | | $\delta_1$ | $\delta_2$ | $\delta_3$ | *abs_rel* | *rmse* | *log_10* | *rmse_log* | *silog* | *sq_rel* |
|---|---|---|---|---|---|---|---|---|---|---|
| KITTI | B | 0.9292 | 0.9883 | 0.9973 | 0.0775 | 3.4291 | 0.0348 | 0.1196 | 10.6605 | 0.3406 |
| | +RE | **0.9433** | **0.9884** | **0.9973** | **0.0640** | **3.2055** | **0.0338** | **0.0982** | **9.5324** | **0.2760** |
| NYU | B | 0.8237 | 0.9597 | 0.9895 | 0.1390 | 0.5006 | 0.0581 | 0.1770 | 15.0716 | 0.1042 |
| | +RE | **0.8634** | **0.9722** | **0.9933** | **0.1205** | **0.4383** | **0.0509** | **0.1556** | **13.1205** | **0.0790** |

Table 8: The model added with redundancy elimination (+RE) outperforms baseline model only using image representations (B) both on KITTI and NYU datasets.

## 5 Conclusion

In this paper, we propose a novel framework to dig out spatial priors, which are embedded in relationships in a scene image, to enhance depth estimation. We strive to answer two questions: (1) which relationships contain useful spatial priors? (2) How much do the spatial priors contribute to the depth estimation? Several modules including 'Spatial alignment', 'Orthogonalization' and 'Attention' are designed to solve the questions. Subsequently, extensive objective and subjective experiments are conducted which strongly proves the proposed model successfully captures spatial priors in relationships. Besides, other complicated MDE models can be inserted to our framework to enhance performance.

**Acknowledgement** This work is in part supported by Natural Sciences and Engineering Research Council of Canada, Canada Research Chair program, NRF Centre for Advanced Robotics Technology Innovation (CARTIN) of Singapore, and AI Singapore (Award No.: AISG2-RP-2021-023). Jiayan Qiu is the corresponding author of this paper.

## References

1. Abu-El-Haija, S., Perozzi, B., Al-Rfou, R., Alemi, A.A.: Watch your step: Learning node embeddings via graph attention. In: Advances in Neural Information Processing Systems. pp. 9180–9190 (2018)
2. Atapour-Abarghouei, A., Breckon, T.P.: Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2800–2810 (2018)
3. Atwood, J., Towsley, D.: Diffusion-convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1993–2001 (2016)
4. Bacciu, D., Errica, F., Micheli, A.: Contextual graph markov model: A deep and generative approach to graph processing. In: ICML (2018)
5. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4009–4018 (2021)
6. Bruna, J., Zaremba, W., Szlam, A., LeCun, Y.: Spectral networks and locally connected networks on graphs. arXiv preprint arXiv:1312.6203 (2013)
7. Cao, Y., Wu, Z., Shen, C.: Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE Transactions on Circuits and Systems for Video Technology **28**(11), 3174–3182 (2017)
8. 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)
9. Chen, J., Zhu, J., Song, L.: Stochastic training of graph convolutional networks with variance reduction. arXiv preprint arXiv:1710.10568 (2017)
10. Chen, P.Y., Liu, A.H., Liu, Y.C., Wang, Y.C.F.: Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2624–2632 (2019)
11. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. Advances in neural information processing systems **29** (2016)
12. Chen, W., Qian, S., Deng, J.: Learning single-image depth from videos using quality assessment networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5604–5613 (2019)
13. Chen, Y.C., Lin, Y.Y., Yang, M.H., Huang, J.B.: Crdoco: Pixel-level domain transfer with cross-domain consistency. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1791–1800 (2019)
14. 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)
15. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: Advances in neural information processing systems. pp. 3844–3852 (2016)
16. Desai, C., Ramanan, D., Fowlkes, C.C.: Discriminative models for multi-class object layout. International journal of computer vision **95**(1), 1–12 (2011)
17. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015)

18. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. arXiv preprint arXiv:1406.2283 (2014)
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. Flynn, J., Neulander, I., Philbin, J., Snavely, N.: Deepstereo: Learning to predict new views from the world's imagery. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5515–5524 (2016)
22. Forsyth, D., Ponce, J.: Computer vision: A modern approach. Prentice hall (2011)
23. 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)
24. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2002–2011 (2018)
25. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Zhang, K., Tao, D.: Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2427–2436 (2019)
26. Fu, H., Li, S., Jia, R., Gong, M., Zhao, B., Tao, D.: Hard example generation by texture synthesis for cross-domain shape similarity learning. Advances in Neural Information Processing Systems **33**, 14675–14687 (2020)
27. Gallicchio, C., Micheli, A.: Graph echo state networks. In: The 2010 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2010)
28. Gao, H., Wang, Z., Ji, S.: Large-scale learnable graph convolutional networks. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1416–1424. ACM (2018)
29. Garg, R., Bg, V.K., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: European conference on computer vision. pp. 740–756. Springer (2016)
30. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
31. Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E.: Neural message passing for quantum chemistry. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1263–1272. JMLR. org (2017)
32. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 270–279 (2017)
33. Gori, M., Monfardini, G., Scarselli, F.: A new model for learning in graph domains. In: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. vol. 2, pp. 729–734. IEEE (2005)
34. 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)
35. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: Advances in Neural Information Processing Systems. pp. 1024–1034 (2017)

36. Handa, A., Whelan, T., McDonald, J., Davison, A.: A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In: IEEE Intl. Conf. on Robotics and Automation, ICRA. Hong Kong, China (May 2014)
37. Henaff, M., Bruna, J., LeCun, Y.: Deep convolutional networks on graph-structured data. arXiv preprint arXiv:1506.05163 (2015)
38. Hoiem, D., Efros, A.A., Hebert, M.: Closing the loop in scene interpretation. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
39. Huang, W., Zhang, T., Rong, Y., Huang, J.: Adaptive sampling towards fast graph representation learning. In: Advances in Neural Information Processing Systems. pp. 4558–4567 (2018)
40. 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)
41. Karsch, K., Liu, C., Kang, S.B.: Depth transfer: Depth extraction from video using non-parametric sampling. IEEE transactions on pattern analysis and machine intelligence **36**(11), 2144–2158 (2014)
42. Kim, S., Park, K., Sohn, K., Lin, S.: Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In: European conference on computer vision. pp. 143–159. Springer (2016)
43. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
44. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
45. 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)
46. 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**(1), 32–73 (2017)
47. Ladicky, L., Shi, J., Pollefeys, M.: Pulling things out of perspective. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 89–96 (2014)
48. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth international conference on 3D vision (3DV). pp. 239–248. IEEE (2016)
49. Lan, L., Wang, X., Zhang, S., Tao, D., Gao, W., Huang, T.S.: Interacting tracklets for multi-object tracking. IEEE Transactions on Image Processing **27**(9), 4585–4597 (2018)
50. Lee, J.H., Han, M.K., Ko, D.W., Suh, I.H.: From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326 (2019)
51. Lee, J.B., Rossi, R., Kong, X.: Graph classification using structural attention. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1666–1674. ACM (2018)
52. Levie, R., Monti, F., Bresson, X., Bronstein, M.M.: Cayleynets: Graph convolutional neural networks with complex rational spectral filters. IEEE Transactions on Signal Processing **67**(1), 97–109 (2018)

53. Li, B., Shen, C., Dai, Y., Van Den Hengel, A., He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1119–1127 (2015)
54. Li, R., Wang, S., Zhu, F., Huang, J.: Adaptive graph convolutional neural networks. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
55. 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)
56. Liang, K., Guo, Y., Chang, H., Chen, X.: Visual relationship detection with deep structural ranking. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
57. 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)
58. Lienen, J., Hullermeier, E., Ewerth, R., Nommensen, N.: Monocular depth estimation via listwise ranking using the plackett-luce model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14595–14604 (2021)
59. 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)
60. Liu, B., Gould, S., Koller, D.: Single image depth estimation from predicted semantic labels. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 1253–1260. IEEE (2010)
61. Liu, B., Dong, Q., Hu, Z.: Zero-shot learning from adversarial feature residual to compact visual feature. In: AAAI. pp. 11547–11554 (2020)
62. Liu, B., Dong, Q., Hu, Z.: Hardness sampling for self-training based transductive zero-shot learning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16499–16508 (2021)
63. Liu, B., Dong, Q., Hu, Z.: Semantic-diversity transfer network for generalized zero-shot learning via inner disagreement based ood detector. Knowledge-Based Systems **229**, 107337 (2021)
64. Liu, B., Hu, L., Hu, Z., Dong, Q.: Hardboost: Boosting zero-shot learning with hard classes. arXiv preprint arXiv:2201.05479 (2022)
65. Liu, H., Yang, Y., Wang, X.: Overcoming catastrophic forgetting in graph neural networks. In: Proceedings of the AAAI conference on artificial intelligence (2021)
66. Liu, Z., Chen, C., Li, L., Zhou, J., Li, X., Song, L., Qi, Y.: Geniepath: Graph neural networks with adaptive receptive paths. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 4424–4431 (2019)
67. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European conference on computer vision. pp. 852–869. Springer (2016)
68. Lu, Y., Pirk, S., Dlabal, J., Brohan, A., Pasad, A., Chen, Z., Casser, V., Angelova, A., Gordon, A.: Taskology: Utilizing task relations at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8700–8709 (2021)
69. Maksai, A., Wang, X., Fleuret, F., Fua, P.: Non-markovian globally consistent multi-object tracking. In: The IEEE International Conference on Computer Vision (ICCV) (2017)

70. Maksai, A., Wang, X., Fua, P.: What players do with the ball: A physically constrained interaction modeling. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
71. Mertan, A., Sahin, Y.H., Duff, D.J., Unal, G.: A new distributional ranking loss with uncertainty: Illustrated in relative depth estimation. In: 2020 International Conference on 3D Vision (3DV). pp. 1079–1088. IEEE (2020)
72. Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., Bronstein, M.M.: Geometric deep learning on graphs and manifolds using mixture model cnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5115–5124 (2017)
73. Niepert, M., Ahmed, M., Kutzkov, K.: Learning convolutional neural networks for graphs. In: International conference on machine learning. pp. 2014–2023 (2016)
74. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32**, 8026–8037 (2019)
75. 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)
76. Qi, X., Liao, R., Liu, Z., Urtasun, R., Jia, J.: Geonet: Geometric neural network for joint depth and surface normal estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 283–291 (2018)
77. 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 (2019)
78. 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)
79. Qiu, J., Yang, Y., Wang, X., Tao, D.: Hallucinating visual instances in total absentia. In: European Conference on Computer Vision (2020)
80. 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)
81. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. arXiv preprint arXiv:1907.01341 (2019)
82. Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., Black, M.J.: Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12240–12249 (2019)
83. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks (2016)
84. Ren, S., Zhou, D., He, S., Feng, J., Wang, X.: Shunted self-attention via multi-scale token aggregation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
85. Ren, Z., Lee, Y.J.: Cross-domain self-supervised multi-task feature learning using synthetic imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 762–771 (2018)
86. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE transactions on neural networks **20**(1), 61–80 (2008)

87. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International journal of computer vision **47**(1), 7–42 (2002)
88. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. pp. 746–760. Springer (2012)
89. Sperduti, A., Starita, A.: Supervised neural networks for the classification of structures. IEEE Transactions on Neural Networks **8**(3), 714–735 (1997)
90. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10781–10790 (2020)
91. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
92. Veličković, P., Fedus, W., Hamilton, W.L., Liò, P., Bengio, Y., Hjelm, R.D.: Deep graph infomax. arXiv preprint arXiv:1809.10341 (2018)
93. Wang, C., Buenaposada, J.M., Zhu, R., Lucey, S.: Learning depth from monocular videos using direct methods. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2022–2030 (2018)
94. Wang, C., Wang, C., Xu, C., Tao, D.: Tag disentangled generative adversarial networks for object image re-rendering. In: International joint conference on artificial intelligence (IJCAI) (2017)
95. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.L.: Towards unified depth and semantic prediction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2800–2809 (2015)
96. Wang, X., Turetken, E., Fleuret, F., Fua, P.: Tracking interacting objects optimally using integer programming. In: European Conference on Computer Vision and Pattern Recognition (ECCV). pp. 17–32 (2014)
97. Wang, X., Turetken, E., Fleuret, F., Fua, P.: Tracking interacting objects using intertwined flows. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(11), 2312–2326 (2016)
98. Xian, K., Shen, C., Cao, Z., Lu, H., Xiao, Y., Li, R., Luo, Z.: Monocular relative depth perception with web stereo data supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 311–320 (2018)
99. Xian, K., Zhang, J., Wang, O., Mai, L., Lin, Z., Cao, Z.: Structure-guided ranking loss for single image depth prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 611–620 (2020)
100. Xu, D., Ouyang, W., Wang, X., Sebe, N.: Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 675–684 (2018)
101. 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)
102. Yang, Y., Feng, Z., Song, M., Wang, X.: Factorizable graph convolutional networks. In: Neural Information Processing Systems (NeurIPS) (2020)
103. Yang, Y., Qiu, J., Song, M., Tao, D., Wang, X.: Distilling knowledge from graph convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
104. Yang, Y., Wang, X., Song, M., Yuan, J., Tao, D.: SPAGAN: shortest path graph attention network. In: International Joint Conference on Artificial Intelligence (IJCAI) (2019)

105. Yang, Z., Liu, D., Wang, C., Yang, J., Tao, D.: Modeling image composition for complex scene generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7764–7773 (2022)
106. 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)
107. Ying, Z., You, J., Morris, C., Ren, X., Hamilton, W., Leskovec, J.: Hierarchical graph representation learning with differentiable pooling. In: Advances in Neural Information Processing Systems. pp. 4800–4810 (2018)
108. Yu, B., Liu, T., Gong, M., Ding, C., Tao, D.: Correcting the triplet selection bias for triplet loss. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 71–87 (2018)
109. Yu, B., Tao, D.: Deep metric learning with tuplet margin loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6490–6499 (2019)
110. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4471–4480 (2019)
111. 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)
112. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
113. 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)
114. Zhang, J., Shi, X., Xie, J., Ma, H., King, I., Yeung, D.Y.: Gaan: Gated attention networks for learning on large and spatiotemporal graphs. arXiv preprint arXiv:1803.07294 (2018)
115. Zhang, Z., Cui, Z., Xu, C., Jie, Z., Li, X., Yang, J.: Joint task-recursive learning for semantic segmentation and depth estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 235–251 (2018)
116. Zhang, Z., Cui, Z., Xu, C., Yan, Y., Sebe, N., Yang, J.: Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4106–4115 (2019)
117. Zhao, S., Fu, H., Gong, M., Tao, D.: Geometry-aware symmetric domain adaptation for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9788–9798 (2019)
118. Zheng, C., Cham, T.J., Cai, J.: T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 767–783 (2018)
119. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1851–1858 (2017)
120. 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)

121. Zoran, D., Isola, P., Krishnan, D., Freeman, W.T.: Learning ordinal relationships for mid-level vision. In: Proceedings of the IEEE international conference on computer vision. pp. 388–396 (2015)