Supplementary Material
Relationformer: A Unified Framework for *Image-to-Graph* Generation

Suprosanna Shit[⋆1,3], Rajat Koner[⋆2], Bastian Wittmann[1], Johannes Paetzold[1], Ivan Ezhov[1], Hongwei Li[1], Jiazhen Pan[1], Sahand Sharifzadeh[2], Georgios Kaissis[1], Volker Tresp[2], and Bjoern Menze[3]

[1] Technical University of Munich, Munich, Germany
[2] Ludwig Maximilian University of Munich, Munich, Germany
[3] University of Zurich, Zurich, Switzerland
suprosanna.shit@tum.de, koner@dbs.ifi.lmu.de

A Transformer and Deformable-DETR

The core of a transformer [12] is the attention mechanism. Let us consider an image feature map $\boldsymbol{f}_I$, the q^{th} query with associated features $\boldsymbol{f}_q$ and k^{th} key with associated image features $\boldsymbol{f}_I^k$. One can define the multi-head attention for M number of heads and K number of key elements as

$$\text{MultiHeadAttn}\left(\boldsymbol{f}_q, \boldsymbol{f}_I\right) = \sum_{m=1}^{M} \boldsymbol{W}_m \left[\sum_{k=1}^{K} \boldsymbol{A}_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{f}_I^k\right]$$

where $\boldsymbol{W}'_m$ and $\boldsymbol{W}_m$ are learnable weights. The attention weights $\boldsymbol{A}_{mqk} \propto \exp\left\{\frac{\boldsymbol{f}_q^\top \boldsymbol{W}''^\top_m \boldsymbol{W}'''_m \boldsymbol{f}_I^k}{\sqrt{d_k}}\right\}$ are normalized as $\sum_{k=1}^{K} \boldsymbol{A}_{mqk} = 1$, where $\boldsymbol{W}''_m, \boldsymbol{W}'''_m$ are also learnable weights and d_k is the temperature parameter. To differentiate position of each element uniquely, $\boldsymbol{f}_q$ and $\boldsymbol{f}_I$ are given a distinct positional embedding.

In our work, we use the multi scale deformable attention [14]. Let us consider the reference point associated with $\boldsymbol{f}_q$ as $\boldsymbol{x}_q$. First, for the m^{th} attention head, we need to compute the k^{th} sampling offset $\Delta\boldsymbol{x}_{mqk}$ based on the query features $\boldsymbol{f}_q$. Subsequently, the sampled image features $\boldsymbol{f}_I(\boldsymbol{x}_q + \Delta\boldsymbol{x}_{mqk})$ go through a single layer $\boldsymbol{W}'_m$ followed by a multiplication with the attention weight $\boldsymbol{A}_{mqk}$, which is also obtained from the query features $\boldsymbol{f}_q$. Finally, another single layer $\boldsymbol{W}_m$ merges all the heads. Formally, the deformable attention operation (DefAttn) for M heads and K sampling points is defined as:

$$\text{DefAttn}(\boldsymbol{f}_q, \boldsymbol{x}_q, \boldsymbol{f}_I) = \sum_{m=1}^{M} \boldsymbol{W}_m \left[\sum_{k=1}^{K} \boldsymbol{A}_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{f}_I(\boldsymbol{x}_q + \Delta\boldsymbol{x}_{mqk})\right] \quad (1)$$

⋆ equal contribution

The multi-scale deformable attention for L number of level is given as

$$\text{MSDefAttn}(\boldsymbol{f}_q, \boldsymbol{x}_q, \{\boldsymbol{f}_I^l\}_{l=1}^L) = \sum_{m=1}^{M} \boldsymbol{W}_m \left[\sum_{l=1}^{L}\sum_{k=1}^{K} \boldsymbol{A}_{mlqk} \cdot \boldsymbol{W}^{'}_m \boldsymbol{f}_I^l(\phi_l(\boldsymbol{x}_q) + \Delta \boldsymbol{x}_{mlqk})\right]$$

where ϕ_l rescales the normalized reference point coordinates appropriately in the corresponding image space.

B Dataset

Here we describe the individual datasets used in our experimentation in detail. We also elaborate on generating train-test sets for our experiments. For 20 U.S. Cities and 3D synthetic vessel we extract overlapping patches from large images. This provides us a large enough sample size to train our Relationformer from scratch. Since, a DETR like architecture is not translation invariant because of learned [obj]-tokens in the decoder, extracting overlapping patches drastically increases the effective sample size within a limited number of available images.

B.1 Toulouse Road Network

The Toulouse Road Network dataset [1] is based on publicly available satellite images from Open Streetmap and consists of semantic segmentation images with their corresponding graph representations. For our experiments we use the same split as in the original dataset paper with 80,357 samples in the training set, 11,679 samples in the validation set, and 18,998 samples in the test set [1].

B.2 20 U.S. Cities Dataset

For the 20 U.S. Cities dataset [3], there are 180 images with a resolution of 2048x2048. We select 144 for training, 9 for validation, and 27 for testing. From those images, we extract overlapping patches of size 128x128 to construct the final train-validation-test split. We crop the RGB image and the corresponding graph followed by a node simplification. Following Belli et al. [1], we prune the dense nodes by computing the angle between two road-segments at each node of degree 2 and only keep a node if the road curvature is less than 160 degrees. This allows eliminating redundant nodes and simplifying the graph prediction task. Fig. 1 illustrates the pruning process.

B.3 3D Synthetic Vessels

Our synthetic vessel dataset is based on publicly available synthetic images generated in Tetteh et al. [11]. In this dataset, the ground truth graph was generated by [10] and from that, corresponding voxel-level semantic segmentation data was generated. Grey valued data was obtained by adding different noise levels to the

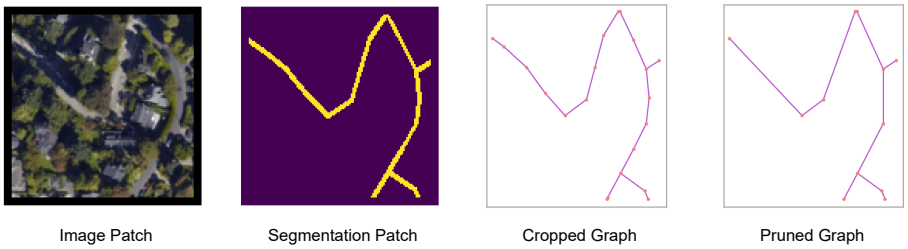

Fig. 1. Preprocessing steps for the 20 U.S. Cities dataset. The same steps are followed in the 3D Synthetic Vessel dataset curation.

segmentation map. Specifically, we train on greyscale "images" and their corresponding vessel graph representations, where each node represents a bifurcation point, and the edges represent their connecting vessels. The whole dataset contains 136 3D volumes of size 325x304x600. First, we choose 40 volumes to create a train and validation set and next pick 10 volumes for the test set. From this, we extract overlapping patches of size 64x64x64 to construct the final train-validation-test set. Similar to the 20 U.S. cities dataset, we prune nodes having degree 2 based on the angle between two edges.

B.4 Visual Genome

Visual Genome is one of the largest scene graph datasets consisting of 108,077 natural images [6]. However, the original dataset suffers from multiple annotation errors and improper bounding boxes. Lu et al. [9] proposed a refined version of Visual Genome with the most frequent occurring 150 objects classes and 50 relation categories. It also proposed its own train/val/test splits and is the most widely used data-split [13,5,7,8] for SGG. For fair comparison, we only train on the Visual Genome dataset and do **not** use any pre-training.

C Metrics Details

Metrics for Spatio-Structural Graph: We use three different kinds of metrics to capture spatial similarity alongside the topological similarity of the predicted graphs. The graph-level metrics include; 1) *Street Mover Distance (SMD):* SMD[1] compute Wasserstein distance between the uniformly sampled fixed number of points (See Fig. 2) from the predicted and ground truth edges; and 2) *TOPO Score:* TOPO Score[3] computes precision, recall, and F-1 score for topological mismatch in terms of the false-positive and false-negative topological loop. Alongside, we use 3) *Node Detection:* For this, we report mean average precision (mAP) and mean average recall (mAR) over a threshold range [0.5,0.95,0.05] for node box prediction. Similarly, we use 4) *Edge Detection:* We compute the mAP and mAR for the edge in the same way as above. The edge

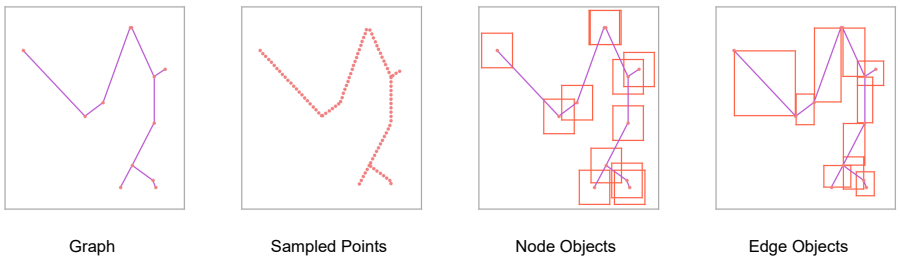

Fig. 2. Sampled points, node objects and edge objects for computing different spatio-structural graph metrics. The same notion is used for 3D graphs.

boxes are constructed from the center points of two connecting nodes (See Fig. 2). For vertical and horizontal edges we assume an hypothetical width of 0.15 to avoid objects with near zero width.

Metrics for Spatio-Semantic Graph: We evaluate Relationformer on the most challenging Scene Graph Detection(SGDet) metrics and its variants. Unlike other scene graph metrics like Predicate Classification (PredCls) or Scene Graph classification (SGCls) , SGDet does not use apriori information on class label or object spatial position and does not rely on complex RoI-align based spatial features. SGDet jointly measures the predicted boxes (with 50% overlaps) class labels of an object, and relation labels. The variants of SGDet include 1) *Recall:* Recall at the different K (20, 50 and 100) of predicted relation that reflects overall relation prediction performance, 2) *Mean-Recall:* mean-Recall computes mean of each relation class-wise recall that reflects the performance under the relational imbalance or long-tailed distribution of relation class, 3) *ng-Recall:* ng-Recall is recall w/o graph constraints on the prediction, which takes the top-k predictions instead of just the top-1. Additionally, we use 4) AP@50: Average precision at 50% threshold of IOU reflects an average object detection performance.

D Model Details

Table 1. The model parameters used in Relationformer experiments across the various datasets. Specifically, we list details on the backbone and the transformer's number of layers, feature dimension and other details.

DataSet	Backbone	Transformer				MLP Dim
		Enc. Layer	Dec. Layer	# [obj]-tokens	d_{emb}	
Toulouse	ResNet-50	4	4	20	256	512
20 US cities	ResNet-101	4	4	80	512	1024
Synth Vessel	SE-Net	4	4	80	256	1024
Visual Genome	ResNet-50	6	6	200	512	2048

Table 1, describes the backbone and important parameters of the Relationformer. We experiment with different ResNet backbones to show the flexibility of our Relationformer. In order to reduce energy consumption, we use the lighter ResNet50 for most 2D datasets. For the 3D experiment, we used Squeeze-and-Excite Net [4]. We used the number of encoder and decoder layers and the number of [obj]-tokens in the increasing order of dataset complexity. We find that four transformer layers and 20 [obj]-tokens suffice for Toulouse, while we need four transformer layers and 80 [obj]-tokens are required for 20 U.S. cities and synthetic vessel datasets. We need 6 layers of transformer and 200 [obj]-tokens for the visual genome. The ablation on the number of transformer layers and number of [obj]-tokens are shown in the next section.

E Training Details

Table 2. A list of the important set of parameters used in Relationformer for respective training. Furthermore, we list the weights for bipartite matching costs and training losses.

DataSet	Batch Size	Learning rate	Epoch	Cost Coeff.			Loss Coeff.			
				cls	reg	gIoU	λ_{reg}	λ_{gIoU}	λ_{cls}	λ_{rln}
Toulouse	64	10^{-4}	50	2	5	0	5	2	2	1
20 US cities	32	10^{-4}	100	3	5	0	5	2	3	4
3D Vessel Net	48	10^{-4}	100	2	5	0	2	3	3	4
Visual Genome	16	10^{-4}	25	3	2	3	2	2	4	6

Table. 2, summarizes some principal parameters we use in the training. We use AdamW optimizer with a step learning rate. For scene graph generation, we use the prior statistical distribution or frequency-bias [13] of relation for each subject-object pair. To minimize the data imbalance for a relation label present in the Visual Genome, we use log-softmax distribution [7] to soften the frequency bias. Finally, we add this distribution with the predicted relation distribution from the relation head. For the spatio-structural dataset, we set the cost coefficient for the GIoU in the bipartite matcher to be zero because we assume 0.2 widths of the normalized box for each node. Hence, ℓ_1 cost is sufficient to consider for the spatial distances.

F More Ablation Studies on [obj]-tokens and Transformer

We conduct two more ablation studies on Visual Genome for analyzing the influence of [obj]-tokens and optimal number of layers in transformer for the joint graph generation. Furthermore Figure. 3 gives additional insight how [rln]-token is beneficial for joint object-relation graph.

Table 3. Impact of the [obj]-tokens on joint object and relation detection.

#[obj]-tokens	AP@50	R@20	R@50	R@100
75	25.1	20.6	26.1	29.5
100	25.8	21.1	27.4	30.6
200(ours)	26.3	22.2	28.4	31.3
300	26.3	21.9	27.9	31.0

Table 4. Impact of the transformer's layers on joint object-relation detection

# layer	AP@50	R@20	R@50	R@100
4	24.6	20.5	26.5	28.8
5	25.2	21.0	27.2	29.9
6(ours)	26.3	22.2	28.4	31.3

As shown in Table 3, it can be observed that increasing [obj]-tokens does increase object and relation detection performance. However, it becomes relatively stable with increasing object quarries. DETR-like architectures rely on an optimal number of [obj]-tokens to balance positive and negative simple which also helps in object detection as observed in [2]. Thus, in a joint object and relation prediction, a gain might come from optimal number [obj]-tokens, as relation prediction is linearly co-related to object detection performance. It demonstrates that joint object and relation detection can perfectly coexist without hurting the object detection performance. Instead, it can exploit [obj]-tokens enriched with global relational reasoning for efficient relation extraction.

During the ablation with transformer layers, we observe decreasing number of transformer layers shows an initial gain in object and relation detection. However, they lead to early plateau and inferior performance as depicted in table 4. One intuitive reason is that with less parameter and insufficient contextualization Relationformer quickly learn the initial biases present in both object and relation detection and failed to learn the complex global scenario. We use the same number of layers for both encoder and decoder.

G Qualitative Results

Fig. 4 and 5 shows additional qualitative example from our experiments.

References

1. Belli, D., Kipf, T.: Image-conditioned graph generation for road network extraction. arXiv preprint arXiv:1910.14388 (2019)
2. Carion, N., et al.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020)
3. He, S., et al.: Sat2Graph: road graph extraction through graph-tensor encoding. In: European Conference on Computer Vision. pp. 51–67. Springer (2020)
4. Hu, J., et al.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
5. Koner, R., et al.: Relation transformer network. arXiv preprint arXiv:2004.06193 (2020)
6. Krishna, R., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. arXiv preprint arXiv:1602.07332 (2016)
7. Lin, X., et al.: 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)
8. Liu, H., et al.: Fully convolutional scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11546–11556 (2021)
9. Lu, C., et al.: Visual relationship detection with language priors. In: European Conference on Computer Vision (2016)
10. Schneider, M., et al.: Tissue metabolism driven arterial tree generation. Med Image Anal. **16**(7), 1397–1414 (2012)
11. Tetteh, G., et al.: Deepvesselnet: Vessel segmentation, centerline prediction, and bifurcation detection in 3-d angiographic volumes. Frontiers in Neuroscience **14**, 1285 (2020)
12. Vaswani, A., et al.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
13. Zellers, R., et al.: Neural motifs: Scene graph parsing with global context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5831–5840 (2018)
14. Zhu, X., et al.: Deformable DETR: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)

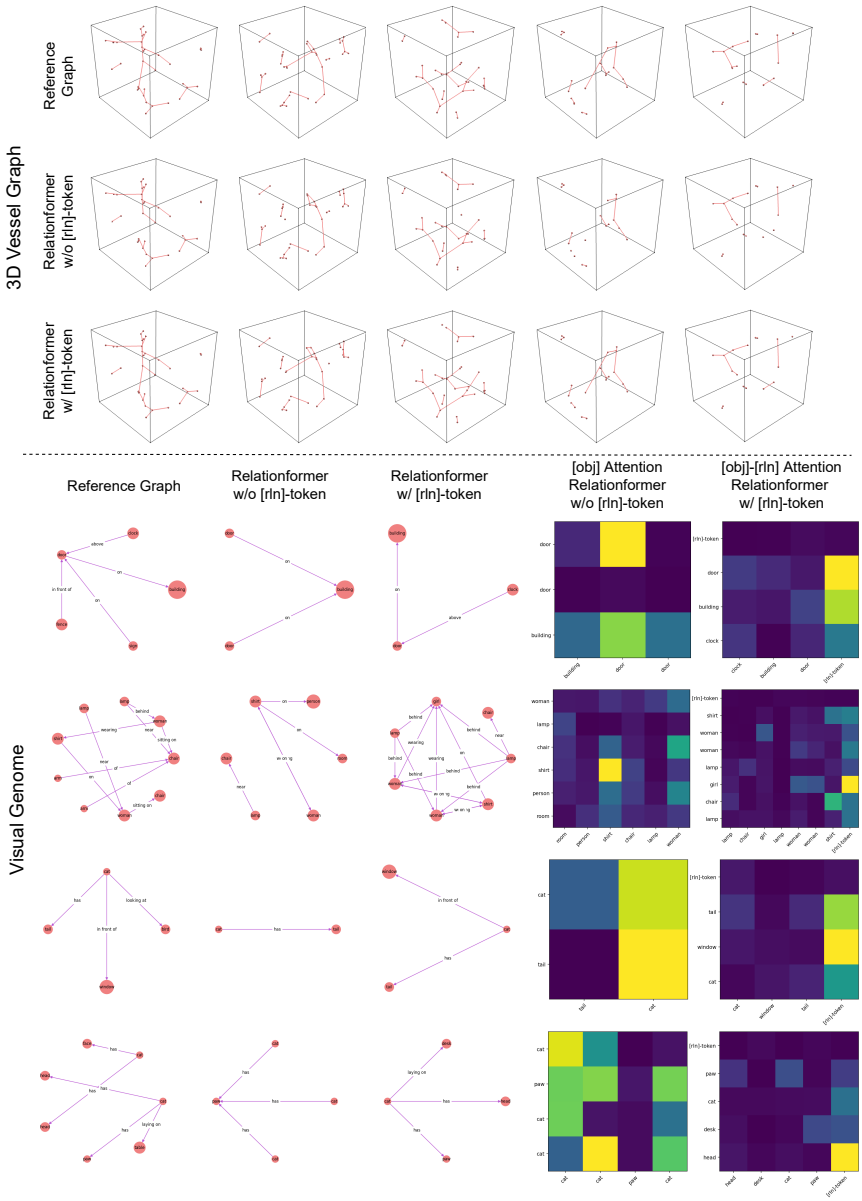

Fig. 3. Typical qualitative results (please zoom in) from our ablation on the synthetic vessel-graph and visual genome datasets. We observe that Relationformer w/o [`rln`]-token is missing vessel edges while Relationformer w/ [`rln`]-token produces correct edges. For visual genome, we can see w/o [`rln`]-token the [`obj`]-tokens have to carry extra burden for relation prediction and sometimes fail to incorporate the global relation. However, the inclusion of [`rln`]-token provides an additional path to flow relation information that benefits the joint object and relation detection.

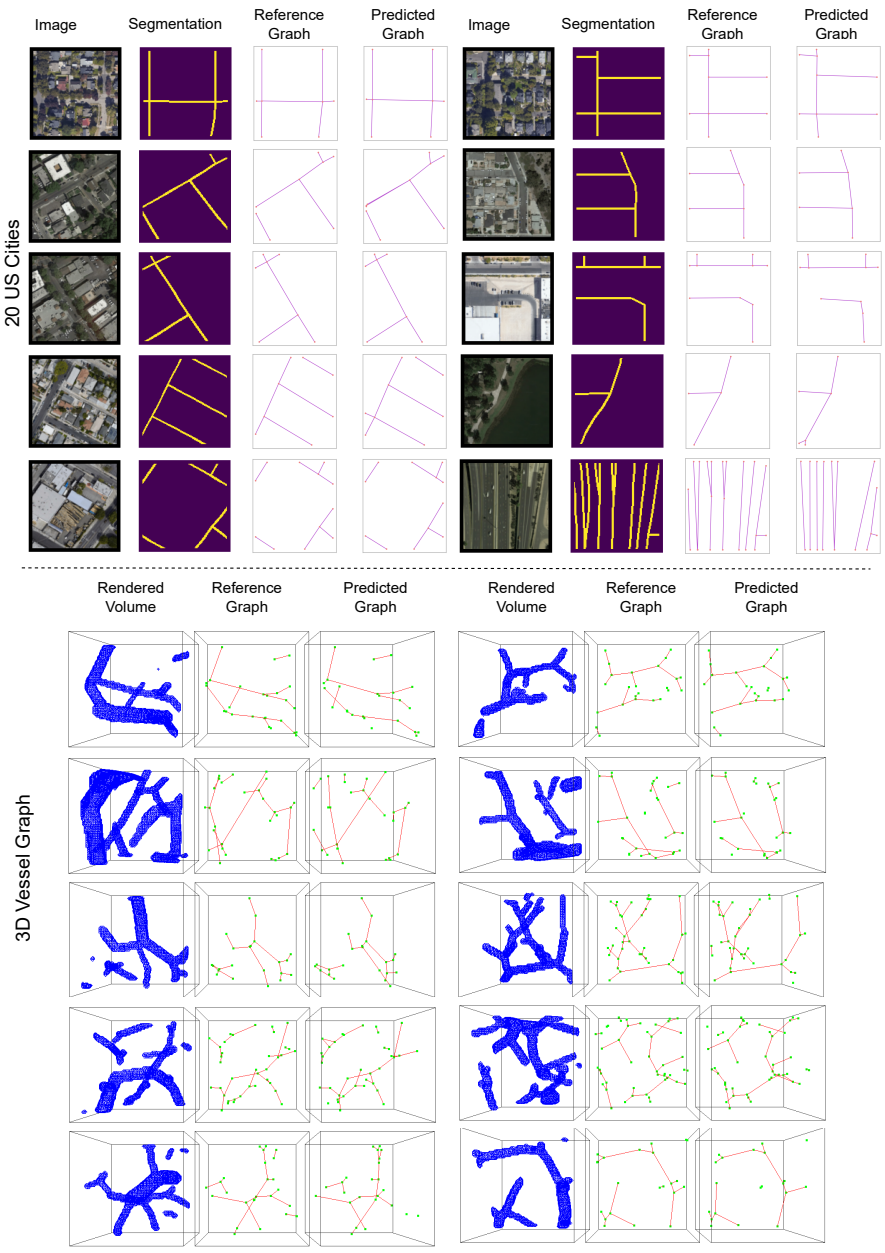

Fig. 4. Qualitative results (please zoom in) for the 20 US cities road-network and synthetic vessel-graph experiments. We observe that Relationformer is able to produce correct results. The segmentation map is given for better interpretability of road network satellite images. For vessel-graphs, we surface-render the segmentation of the corresponding greyscale voxel data.

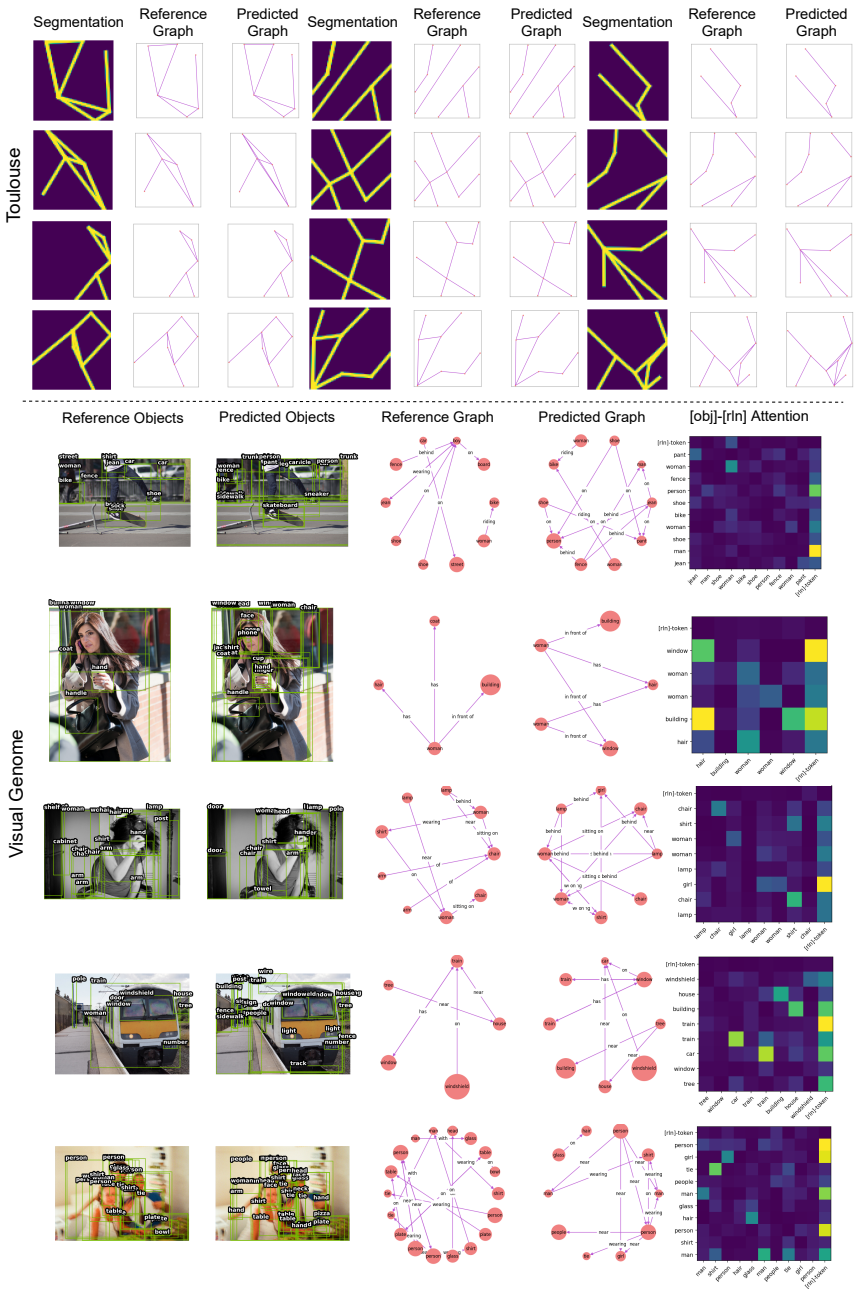

Fig. 5. Qualitative results (please zoom in) from the Toulouse road-network and scene-graph generation experiments. For both datasets, we observe that Relationformer is able to generate an accurate graph. For scene graphs, we visualize the attention map between detected [`obj`]-tokens and [`rln`]-token, which shows that the [`rln`]-token actively attends to objects that contribute to relation formation.