Method	Horse→Zebra		Label→Cityscape		Map→Satellite		Summer→Winter		Apple→Orange	
	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓
CycleGAN (**?**)	77.2	1.957	76.3	3.532	54.6	3.430	84.9	1.022	174.6	10.051
MUNIT (**?**)	133.8	3.790	91.4	6.401	181.7	12.03	115.4	4.901	207.0	12.853
Distance (**?**)	72.0	1.856	81.8	4.410	98.1	5.789	97.2	2.843	181.9	11.362
GCGAN (**?**)	86.7	2.051	105.2	6.824	79.4	5.153	97.5	2.755	178.4	10.828
CUT (**?**)	45.5	0.541	56.4	1.611	56.1	3.301	84.3	1.207	171.5	9.642
NEGCUT (**?**)	39.6	0.477	48.5	1.432	51.0	2.338	82.7	1.352	154.1	7.876
LSeSIM (**?**)	38.0	0.422	49.7	2.867	52.4	3.205	83.9	1.230	168.6	10.386
HnegSRC (**?**)	**34.4**	0.438	**46.4**	0.662	49.2	2.531	81.8	1.181	158.3	8.434
Ours	34.5	**0.271**	46.8	**0.605**	**45.9**	**2.112**	**75.8**	**0.845**	**139.1**	**7.134**

Table 1: Quantitative results. Our model outperforms the baselines in both of FID and KID×100 metrics.

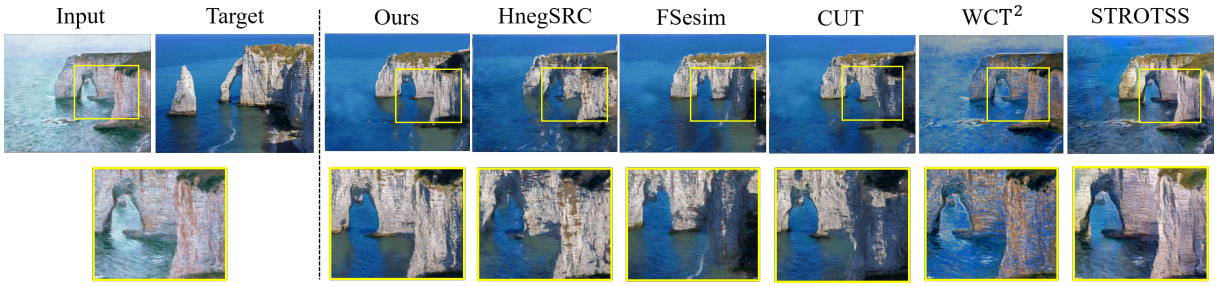

Figure 8: Qualitative comparison on single image translation.

tic image with the enhanced correspondence to the input.

Discussion

The proposed method consists of two parts. First, we construct the graph by the pretrained encoder. Second, we utilize top-K pooling by the pooling vector p to focus on task-relevant nodes which provides the localized graph. To investigate the effectiveness of each part, we first investigate what the vector p learns for the graph pooling procedure. Second, we investigate the adjacency matrix A constructed as in Fig. 10 to verify the patch-wise connection used to construct the graph.

Semantic meaning of the pooling vector p Recall that the vector p allocates higher weights to focus on the important nodes of the graph, which is analogous to the attention mechanism. Here, we provide empirical results which indicates how the vector p allocates weights for nodes Z, V.

Specifically, we visualize $\sigma(S_{in}), \sigma(S_{out})$, given by:

$$S_{in} = p^\top Z \tag{12}$$

$$S_{out} = p^\top V \tag{13}$$

where the σ is sigmoid function. From the result in Fig. 9(b), we can derive two main points. First, the vector p focuses mainly on the object patches which are semantically close and task-relevant. Considering that the top K nodes are selected in graph pooling, the result verifies that the vector p provides focused view of graph by selecting informative nodes. Second, we can observe that the focused parts in $\sigma(S_{in}), \sigma(S_{out})$ are similar. Therefore, the node features Z, V are semantically coherent, indicating the correspondence between the input and the output images.

Adjacency matrix A As shown in Fig. 10, we construct the graph by the learnable adjacency matrix A obtained from the feature F_i, which is the output of the learnable layer h as shown in Fig. 4. We visualize the eigenvectors of the graph Laplacian matrix to verify the learned patch-wise connection in the graph, as suggested in (**?**).

Fig. 9(c) shows that the eigenvectors are semantically coherent with the input image, which clearly demonstrates that the adjacency matrix captures the appropriate implicit semantic connection of the given image.

Ablation study for graphs Here, we provide the ablation study on the graph, such as the number of hops, number of graph pooling layer, value for similarity threshold, and downsampling ratio of the pooling. First, we provide the ablation study for the number of hops and the similarity threshold for A. For the lower ($n = 1$) and larger number of hops ($n = 3$), we observe that the results are degraded. Also, for both the lowered and increased thresholds ($t = 0.0, 0.4, 0.6$), the results are also degraded from the best setting. Especially in the increased threshold (i.e. sparse connectivity), the model shows much degraded performance. This suggests that a sufficiently dense graph can capture the semantically meaningful topology.

Second, we trained the model with different settings for pooling layers. Without the pooling layer (# of pool=0), the performance degraded as the network do not leverage the information from the focused view. For more pooling layers, the model also shows degraded performance, as the pooled

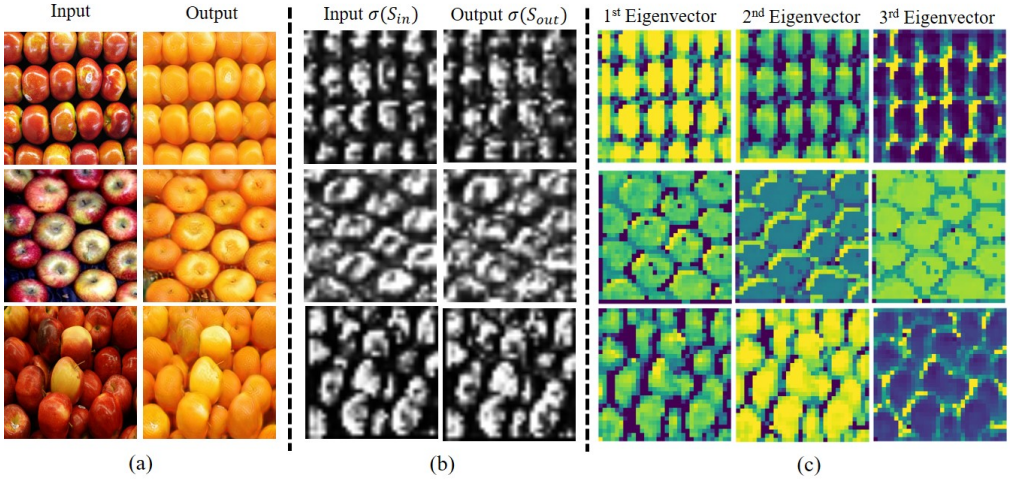

Figure 9: Analysis of the proposed method: (a) Input and the output images. (b) Visualization of $\sigma(S_{in}), \sigma(S_{out})$. The vector p allocates higher weights for the object parts which are task-relevant. Similar appearance refers the correspondence between input and output. (c) Eigenvectors of the Laplacian matrix of A, which are coherent to the semantics of the image.

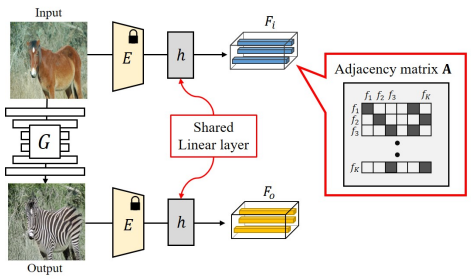

Figure 10: The adjacency matrix A is constructed from F_i which is the output of learnable h. Here, h is updated by the gradient from the F_o similar to CUT (**?**).

	Settings				H→Z	
	# of Hop (**n**)	Thresh (**t**)	# of Pool	Down sample	FID↓	KID↓
GNN Ablation (**n, t**)	1	0.1	1	1/4	37.9	0.438
	3	0.1	1	1/4	39.9	0.374
	2	0.0	1	1/4	34.5	0.551
	2	0.4	1	1/4	36.8	0.293
	2	0.6	1	1/4	38.3	0.332
Pooling Ablation	2	0.1	0	-	37.6	0.432
	2	0.1	2	1/4	35.0	0.625
	2	0.1	1	1/8	37.7	0.340
Proposed	**2**	**0.1**	**1**	**1/4**	**34.5**	**0.271**

Table 2: Quantitative results of ablation studies. Our setting shows the best performance in both of FID and KID×100.

graph has fewer nodes which leads to fewer negative pairs for the contrastive learning. Additionally, we provide the results with varying downsampling rate. For the downsampling of 1/8, the pooled graph consists of fewer nodes, which leads to similar problem with the excessive pooling layers. This again confirms that a sufficiently dense graph after the pooling can capture the semantically meaningful hiearchy. We provide additional ablation study for the graph construction in the supplementary material.

Conclusion

In conclusion, we proposed a novel patch-wise graph representation matching method for image translation task. For structural consistency between input and output images, we proposed to match the constructed graphs between input and outputs. In this part, we used the same adjacency matrix for input and output images for graph consistency. To further leverage the topological information in an hierarchical manner, we applied graph pooling on initial graphs. Our experimental results showed state-of-the-art performance, which again confirms that graph-based patch representation have obvious advantage over baseline methods.

Acknowledgements

This research was supported by National Research foundation of Korea(NRF) (**RS-2023-00262527**)

Ethical Impacts

Regarding on the social impact, the realistic fake images generated by the proposed method may produce a social disinformation, as most of image generation methods shares. Also, the model has potential risk of violating copyright as the model learns the mapping function from input to target distribution. Magnam non mollitia, doloremque autem temporibus non fuga in dolorem nam doloribus praesentium laudantium.Expedita recusandae aliquam unde voluptatum ratione commodi provident, ad eveniet nulla doloremque harum assumenda suscipit, eum fugiat adipisci necessitatibus.