Patch-wise Graph Contrastive Learning for Image Translation

Chanyong Jung[1], **Gihyun Kwon**[1], **Jong Chul Ye**[1, 2]

[1] Department of Brain and Bio Engineering, KAIST, Daejeon, Republic of Korea
[2] Kim Jaechul Graduate School of AI, KAIST, Daejeon, Republic of Korea
{ jcy132, cyclomon, jong.ye }@kaist.ac.kr

Abstract

Recently, patch-wise contrastive learning is drawing attention for the image translation by exploring the semantic correspondence between the input and output images. To further explore the patch-wise topology for high-level semantic understanding, here we exploit the graph neural network to capture the topology-aware features. Specifically, we construct the graph based on the patch-wise similarity from a pretrained encoder, whose adjacency matrix is shared to enhance the consistency of patch-wise relation between the input and the output. Then, we obtain the node feature from the graph neural network, and enhance the correspondence between the nodes by increasing mutual information using the contrastive loss. In order to capture the hierarchical semantic structure, we further propose the graph pooling. Experimental results demonstrate the state-of-art results for the image translation thanks to the semantic encoding by the constructed graphs.

Introduction

Image-to-image translation task is a conditional image generation task in which the model converts the input image into target domain while preserving the content structure of the given input image. The seminar works of image translation models used paired training setting (**?**), or cycle-consistency training (**?**) for content preservation. However, the models have disadvantages in that they require paired dataset or need complex training procedure with additional networks. To overcome the problems, later works introduced one-sided image translation by removing the cycle-consistency (**??**).

Recently, inspired by the success of contrastive learning strategies, Contrastive Unpaired Translation (CUT) (**?**) is proposed to enhance the correspondence between the input and the output images by the patch-wise contrastive learning. The patch-wise contrastive learning is further improved by exploring patch-wise relation such as adversarial hard negative samples (**?**), patch-wise similarity map (**?**), or consistency regularization combined with hard negative mining by patch-wise relation (**?**). Although these methods show meaningful improvement in the performance, they still have a limitation in that the previous works focused only on the individual point-wise matching for each pair, which does not consider the topology with the neighbors (**?**).

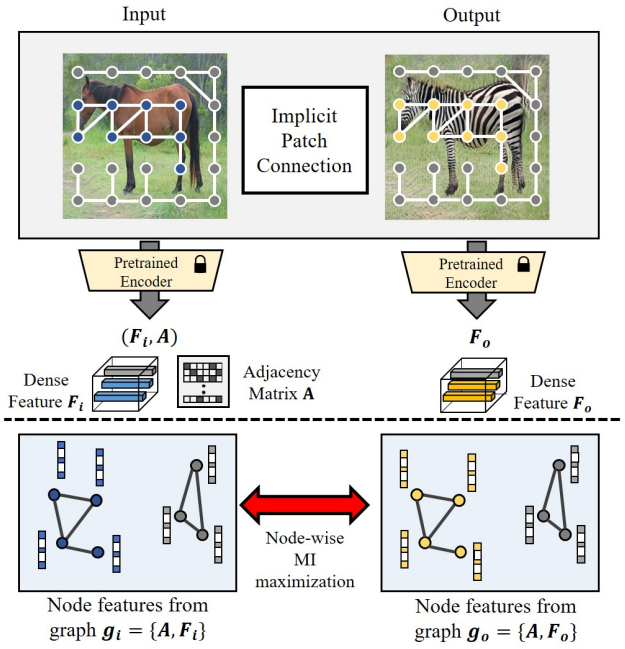

Figure 1: The semantic connectivity of input is extracted by the encoder, and shared to construct the graph network. We maximize the mutual information between the nodes.

To further explore the semantic relationship between the patches, this paper considers image translation tasks as topology-aware representation learning as shown in Fig. 1. Specifically, we propose a novel framework based on the patch-wise graph constrastive learning using the Graph Neural Network (GNN) which is commonly used to extract the feature considering the topological structure.

Several existing works have utilized GNN to capture topology-aware features for various tasks. Hierarchical representation with graph partitioning is proposed for the unsupervised segmentation (**??**), and topology-aware representations (**?**) are extracted based on semantic connectivity between image regions. For knowledge distillation, claimed the *holistic knowledge* (**?**) between the data points is claimed, verifying its effectiveness to encode the topological knowledge of the teacher model.

Despite the great performance in various vision tasks, none of researches have explored the topology-aware features considering the implicit patch-wise semantic connection for the image-to-image translation tasks. Accordingly, here we employ GNN to utilize the patch-wise connection of input image as a prior knowledge for patch-wise contrastive learning. Specifically, we use a pre-trained network to extract the patch-wise features for the input and the output images. Then, we obtain the adjacency matrix calculated by the semantic relation between the patches of the input image, and share it for output image graph. We construct two graphs for the input and the output by the adjacency matrix and the patch features, and obtain the node features by the graph convolution. By maximize the mutual information (MI) between the nodes of input graph and output graph through the contrastive loss, we can enhance the correspondence of patches for the image translation task. Furthermore, to extract the semantic correspondence in a hierarchical manner, we propose to use the graph pooling technique that resembles the attention mechanism.

Our contributions can be summarized as follows:

- We propose a GNN-based framework to capture topology-aware semantic representation by exploiting the patch-wise consistency between the input and translated output images.
- We suggest a method to share the adjacency matrix in order to utilize the patch-wise connection of input image as a prior knowledge for the contrastive learning.
- To further exploit the hierarchical semantic relationship, we propose to use the graph pooling which provides a focused view for the graph.
- Experimental results in five different datasets demonstrates the state-of-the-art performance by producing semantically meaningful graphs.

Related works

Patch-wise contrastive learning for images In a patch-level view, the image has diverse local semantics. The relational knowledge between the patches embodies the correlation between each region, and is utilized for various image generation tasks.

For example, patch-wise contrastive relation (**??**) is utilized for the image translation. Similarly, patch similarity map obtained from pretrained encoder (**?**) is suggested. Recently, patch-level self-correlation map (**?**), query selection module based on patch-wise similarity (**?**), optimal transport plan by patch-wise cost matrix (**?**) are suggested. Also, semantic relation consistency (**?**) is proposed for the image translation tasks. Especially, for style transfer, patch-level relation extracted by vision transformer is recently proposed (**??**). The methods utilized the relation between image tokens to preserve the regional correspondence. Recently, the consistency of the patch-wise semantic relation between the input and the output images was exploited to further improve the correspondence between the input and the output image (**?**). For style transfer, the consistency of patch-level relation extracted by vision transformer was also studied (**??**).

Graph neural network Graph neural network(GNN) learns the representation considering the connectivity of a graph-structured data (**??**). Each node feature models the individual data and its relation to the other data points, aggregating the information from the neighbor nodes.

Thanks to the successes of the GNN to capture the topology-aware features (**???**), the GNN is actively used in various computer vision tasks. For example, the GNN is utilized to capture the local features to find image correspondence (**?**), and multi-modal feature for action segmentation in videos (**?**). Especially, knowledge distillation method through GNN (**??**) is proposed, which is claimed better than conventional contrastive loss, by transferring an additional knowledge on the instance-wise relations.

Recently, the graph constructed by the patch-wise relation was suggested to capture the visual features. The graph partitioning methods are employed for the unsupervised segmentation (**??**), where the graph is obtained by the token-wise similarity from the vision transformer. Vision GNN (**?**) is introduced, which have GCN-based architecture to extract the topology-aware representation, and showed its superior performance to the widely used models such as the CNN and the vision transformers.

Method

Inspired by the previous works, we are interested in exploiting patch-wise relation that represents semantic topology of the image. In particular, we focus on the topology-aware features using graph formed by the semantic relation of patches, and explore how the features improve the task performance.

Specifically, our method is motivated by the consistency of the patch-wise semantic connection of the input and the output images, as shown in Fig. 2(b). If the patch features (z_i, z_j) have semantic connection in the input image, then the patches (v_i, v_j) for the corresponding location of the output should also have the connection. From the motivation, we present a method that utilizes the topology of patch-wise connection of the input image as a prior knowledge.

More specifically, we capture the topology-aware patch features by a GNN, where the patch-wise connection is given by the shared adjacency matrix A. We then obtain the node features $Z = \{z_i\}_{i=1}^N$ and $V = \{v_i\}_{i=1}^N$ and maximize node-wise MI by the contrastive loss. We also utilize the graph pooling, to maximize the MI within the task-relevant focused view of the graph. More details follows.

Graph representation for image translation We first construct the graph for input image $g_i = \{A, F_i\}$, where A is adjacency matrix and F_i's are node features that represent the image patches. Specifically, we randomly sample N patch features $f_n \in \mathbb{R}^c$ from the dense feature $F = E(x) \in \mathbb{R}^{c \times h \times w}$ which is obtained from the intermediate layer of model E, where c, h, w denote the number of color channel, height, and width, respectively. We set the N features as the nodes for the graph g_i (i.e. $F_i = [f_1, , f_N]$).

Then, we obtain the adjacency matrix $A \in \mathbb{R}^{N \times N}$ according to the cosine similarity of the patch features. We connect the patches if the similarity is above the predefined threshold t, and disconnect them in otherwise. Specifically,

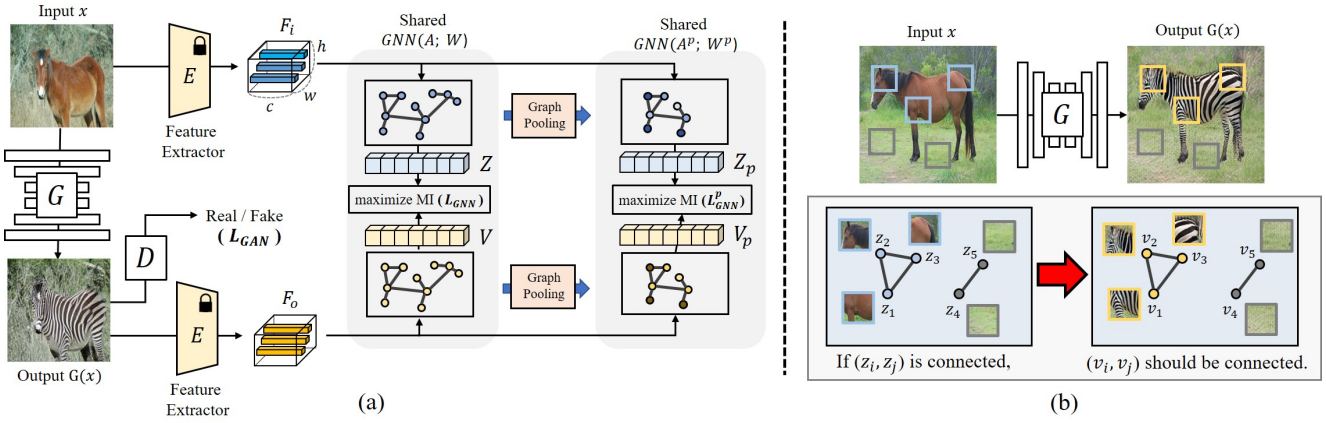

Figure 2: (a) Overall framework of the proposed method. We impose patch-wise regularization by the GNN constructed by the encoder E. We extract the node feature Z, V and maximize $I(Z;V)$. Pooled graphs are utilized to focus on task-relevant nodes. (b) The motivation of the proposed approach to use patch-wise connection of input image as the prior knowledge.

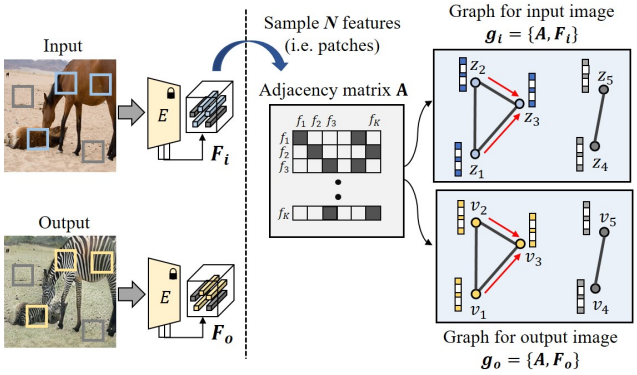

Figure 3: The construction of graphs g_o, g_i with shared adjacency matrix A. Each graph extracts l-hop features Z, V from the given node F_i, F_o.

the connectivity A_{ij} for features f_i, f_j is computed by

$$A_{ij} := \begin{cases} 1 & \text{if } \cos(f_i, f_j) \geq t \\ 0 & \text{if } \cos(f_i, f_j) < t \end{cases} \tag{1}$$

We construct the output graph $g_o = \{A, F_o\}$ in similar way. We sample N features $f'_n \in \mathbb{R}^c$ from the corresponding location of the dense feature $F = E \circ G(x) \in \mathbb{R}^{c \times h \times w}$, and set as the nodes for the graph g_o (i.e. $F_o = [f'_1, , f'_N]$). To retain the topological correspondency between the patches, the output graph inherits the adjacency matrix A from the input graph as shown in Fig. 3.

Next, we obtain the graph representation Z, V using Topology Adaptive Graph Convolution Network (**?**) by the graph g_o, g_i as follows:

$$Z = \sum_{l=0}^{L} (\bar{A})^l F_i W_l \tag{2}$$

$$V = \sum_{l=0}^{L} (\bar{A})^l F_o W_l \tag{3}$$

where $\bar{A}$ is the normalized adjacency matrix, and W_l is the shared parameter for the l-th hop. We obtain 2-hop representation from the graph (i.e. $L = 2$).

Finally, to enforce the topological correspondence between input X and output $G(X)$ for a given generator G, we maximize the mutual information between the nodes Z, V by the infoNCE loss (**?**) as follows:

$$L_{GNN}(X, G(X)) = -\frac{1}{N} \sum_{i=1}^{N} \left[\log \frac{\exp(z_i^\top v_i)}{\sum_{j=1}^{N} \exp(z_i^\top v_j)} \right] \tag{4}$$

where z_i, v_i are the i-th node features from Z and V from X and $G(X)$, respectively.

When $L = 0$, the proposed method shrinks to the conventional patch-wise contrastive learning with the projector network W_0. In this perspective, our method utilizes the higher-ordered features by the graph aggregation (i.e. $L > 0$), which generalizes the conventional contrastive learning.

Graph pooling for focused attention We pool the graph nodes to utilize task-relevant focused attention of the graph. In other words, we downsample the nodes by its relevancy to the task, and construct the graph with fewer nodes to focus on the task-relevant nodes.

Specifically, following the top-K pooling (**?**), we select K nodes from the N nodes $Z = [z_1, , z_N]$ by the similarity score $s_i = p^\top z_i$, where p is the learnable pooling vector. Accordingly, the adjacency matrix $A_p \in \mathbb{R}^{K \times K}$ for the pooled graph is constructed, by excluding the connections with non-selected nodes from the original matrix A. Then, the nodes

are weighted by the score followed by sigmoid funcion σ as:

$$Z_{p,in} = \sigma(S)Z \tag{5}$$

$$V_{p,in} = \sigma(S)V \tag{6}$$

which becomes the input nodes for the pooled graphs. Then, the L-hop features are obtained as:

$$Z_p = \sum_{l=0}^{L} (\bar{A}_p)^l Z_{p,in} W_{p,l} \tag{7}$$

$$V_p = \sum_{l=0}^{L} (\bar{A}_p)^l V_{p,in} W_{p,l} \tag{8}$$

where $W_{p,l}$ is the parameter of the pooled GNN. By constructing the pooled graphs $g_i^p = \{A_p, Z_p\}$, $g_o^p = \{A_p, V_p\}$ and obtaining the l-hop node feature, we also employ the infoNCE loss to maximize the MI between the nodes in the pooled graph as follows:

$$L_{GNN}^{p}(X, G(X)) = -\frac{1}{K}\sum_{i=1}^{K}\left[\log \frac{\exp(z_{p,i}\top v_{p,i})}{\sum_{j=1}^{N}\exp(z_{p,i}\top v_{p,j})}\right] \tag{9}$$

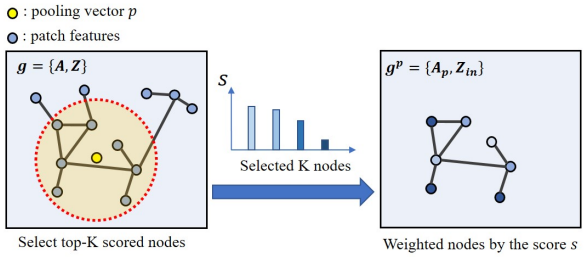

Figure 4: The top-K graph pooling (**?**). The pooling vector p provides the focused view of the graph for the given task. The final node feature is also weighted by p.

Here, it is remarkable how the graph pooling contributes to the improvement. As shown in Fig. 4, the vector p learns to focus on the important nodes, which is determined by the task-relevancy of the nodes. It is analogous to the conventional attention methods (**??**) shown in Fig. 5. Therefore, the graph pooling can be viewed as the node-wise attention, which imposes more regularization for the important nodes to enhance the correspondence for the image translation task.

Overall loss function Our method is one-sided image translation model without cycle-consistency, inspired by the related works based on the patch-wise contrastive learning (**????**). Specifically, the overall loss is given as follows:

$$L_{total} = L_{GAN}(G, D) + \lambda_g \sum_{p=0}^{P} L_{GNN}^{p}(X, G(X)) \tag{10}$$

$$+ \lambda_g \sum_{p=0}^{P} L_{GNN}^{p}(Y, G(Y))$$

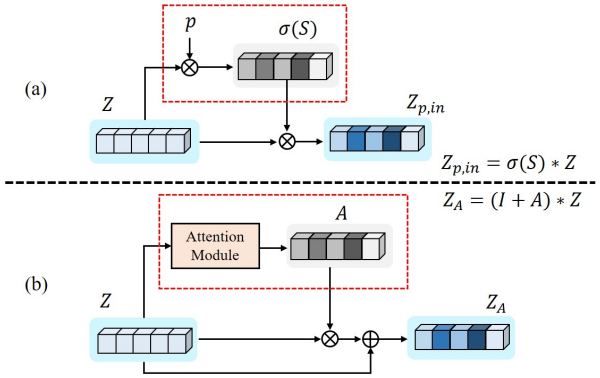

Figure 5: Top-K graph pooling allocates higher weights to the informative nodes, similarly to the attention mechanism. (a) Top-K graph pooling. (b) Attention method.

with generator G and discriminator D shown in Fig. 2(a). L_{GAN} is LSGAN loss (**?**) given as:

$$L_{GAN} = E_{y\sim p_Y}\left[||D(y)||_2^2\right] + E_{x\sim p_X}\left[||1 - D(G(x))||_2^2\right] \tag{11}$$

with the distributions p_X, p_Y for source and target domain. Additionally, we utilize the identity term $L_{GNN}^{p}(Y, G(Y))$ to stabilize the training using the target domain images Y, as suggested in (**?**). $L_{GNN}^{p=0}$ refers the graph loss without the pooling.

Experimental Results

Implementation Details We first verify our method for unpaired image translation task. We verify our method using the five datasets as follows: horse→zebra, Label→Cityscape, map→satellite, summer→winter, and apple→orange. All images are resized into 256×256 for training and testing. Then, we also present our method for single image translation with high resolution, following the previous work (**?**).

For the graph construction, we randomly sampled 256 different patches from the pre-trained VGG16 (**?**) network in both of input and output images. We extract the dense feature from the three different layers (relu3-1, relu4-1, relu4-3layer) inside of the network. For the graph operation, we set the number of GNN hops as 2, and pooling number as 1. For the graph pooling, we downsampled nodes by 1/4. In other words, we have 256 nodes in the initial graph, and 64 nodes for the pooled graph. More details are provided in the supplementary materials.

Image-to-Image translation We compare our method with the two-sided domain translation models, CycleGAN (**?**) and MUNIT (**?**). Also, we selected the one-sided image translation models, DistanceGAN (**?**) and GcGAN (**?**). Especially, since our method is based on the patch-wise contrastive learning, we present the comparison with the recent contrastive learning based methods. We compare our method with CUT (**?**) as baseline model, and the improved model of NEGCUT (**?**), SeSim (**?**) and Hneg-SRC (**?**).

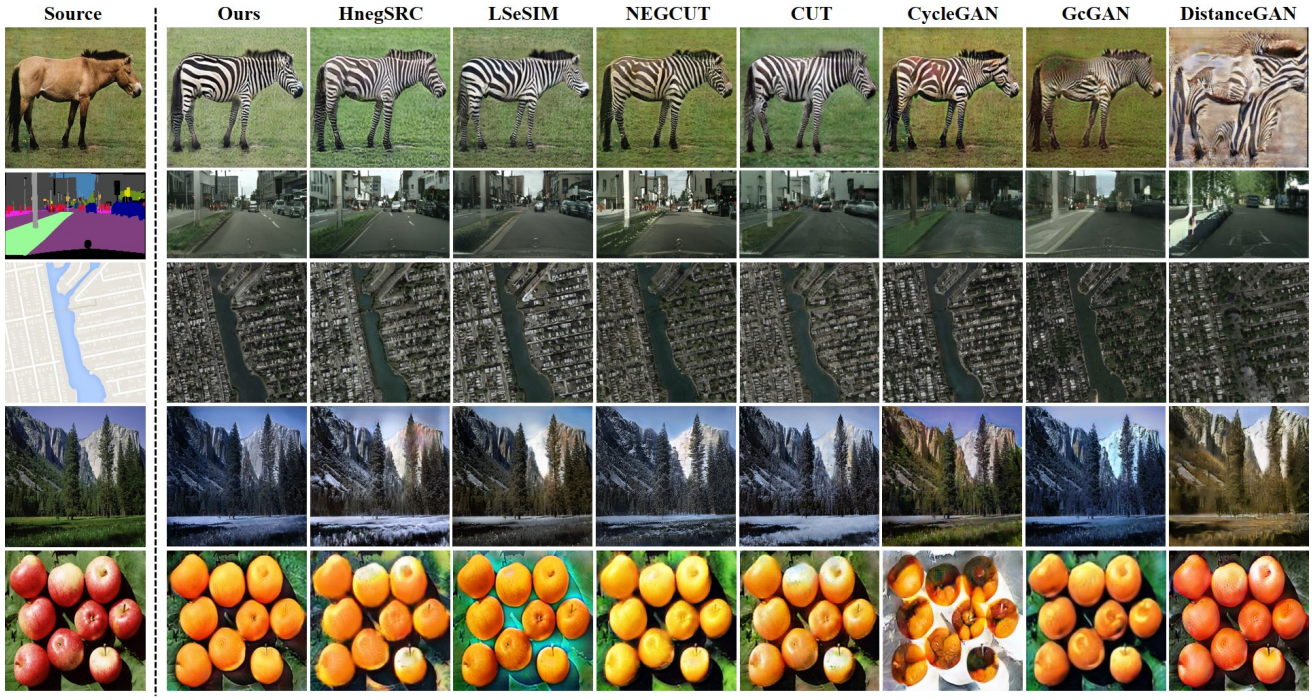

Figure 6: Qualitative comparison with related methods. Our result shows enhanced input-output correspondence, compared to the previous methods.

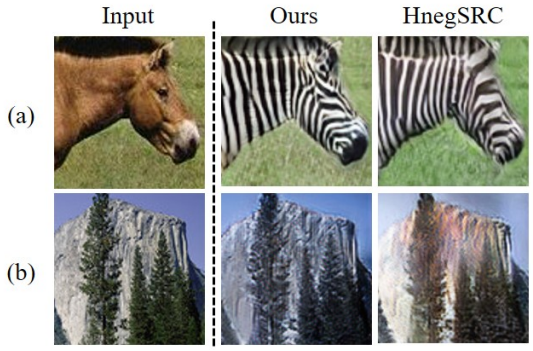

Figure 7: Closer views of the output images. Our method enhances the spatial-specific information given in the input.

Results The results in Fig. 6 verifies that the proposed method generates the images with better visual quality than the other methods, by enhancing the correspondence between the input and the output images. Compared to the other methods, our methods preserves the structural information of the input images, by using the patch-wise connection of the input as the prior knowledge.

Moreover, we further compare our method with the HnegSRC which also utilizes the patch-wise semantic relation of the input. As shown in Fig. 7, our method enhances the spatial-specific information considering the patch-wise semantic neighborhood by the graph operation, compared with the HnegSRC which only imposed the consistency regularization for the patch-wise similarity. Specifically, our method in Fig. 7(a) outputs more realistic zebra by showing the spatial-specific patterns(e.g. dark colored mouth), which is not in the compared result. Also, our result in Fig. 7(b) shows the tree branches with the coherent shapes to the input, which are distorted in the compared method.

The results in Table 1 also supports the outperformance of the proposed method. Specifically, in horse→zebra and Label→Cityscape datasets, we similar FID scores with the HnegSRC, but higher scores by KID. For summer→winter and apple→orange datasets, our model outperformed the others by large margins, which demonstrates the effectiveness of the proposed model.

Single-Image Translation Following the previous work (**?**), we verify our method for the single image translation. The input is a Claude Monet's painting, and the target domain is the natural landscape images. Detailed experimental settings are provided in the supplementary material. For comparison, we choose previous single-image translation models, STROTSS (**?**) and WCT2 (**?**). Also, we selected the patch-based contrastive learning methods, which are CUT (**?**), FSeSim (**?**) and HnegSRC (**?**).

Fig. 8 shows the qualitative comparison of the single-image translation. To show the detailed visual comparison, we enlarge a specific region which is annotated as the yellow box. In WCT2 and STROTSS, the outputs are not fully changed, which contains the artistic textures of the input. The contrastive learning based methods output the improved results, however, show some deformation in shapes. Compared to the previous methods, our method generates realis-

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.

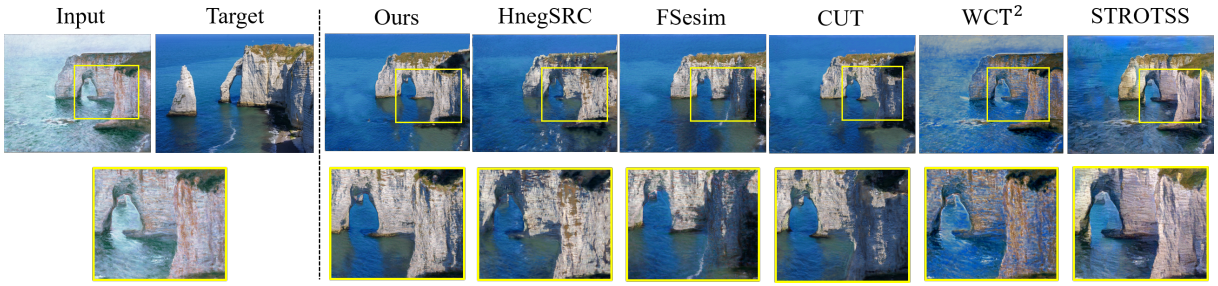

Figure 8: Qualitative comparison on single image translation.

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

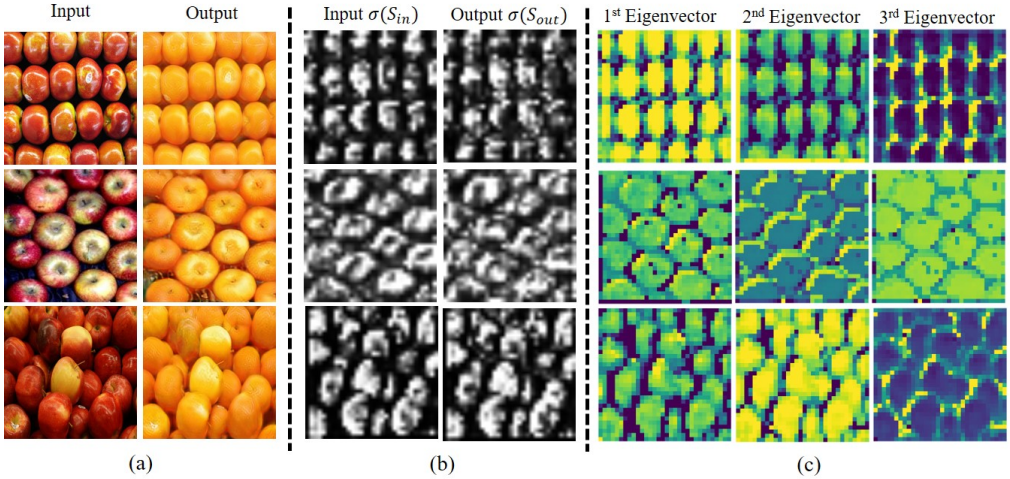

Figure 9: Analysis of the proposed method: (a) Input and the output images. (b) Visualization of $\sigma(S_{in}), \sigma(S_{out})$. The vector p allocates higher weights for the object parts which are task-relevant. Similar appearance refers the correspondence between input and output. (c) Eigenvectors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

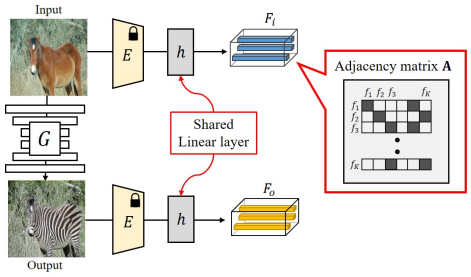

Figure 10: The adjacency matrix A is constructed from F_i which is the output of learnable h. Here, h is updated by the gradient from the F_o similar to CUT (**?**).

	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.