QS-Attn: Query-Selected Attention for Contrastive Learning in I2I Translation

Xueqi Hu[1], Xinyue Zhou[1], Qiusheng Huang[1], Zhengyi Shi[1], Li Sun[1,2]*, Qingli Li[1]

[1]Shanghai Key Laboratory of Multidimensional Information Processing,
[2]Key Laboratory of Advanced Theory and Application in Statistics and Data Science,
East China Normal University, Shanghai, China

Abstract

Unpaired image-to-image (I2I) translation often requires to maximize the mutual information between the source and the translated images across different domains, which is critical for the generator to keep the source content and prevent it from unnecessary modifications. The self-supervised contrastive learning has already been successfully applied in the I2I. By constraining features from the same location to be closer than those from different ones, it implicitly ensures the result to take content from the source. However, previous work uses the features from random locations to impose the constraint, which may not be appropriate since some locations contain less information of source domain. Moreover, the feature itself does not reflect the relation with others. This paper deals with these problems by intentionally selecting significant anchor points for contrastive learning. We design a query-selected attention (QS-Attn) module, which compares feature distances in the source domain, giving an attention matrix with a probability distribution in each row. Then we select queries according to their measurement of significance, computed from the distribution. The selected ones are regarded as anchors for contrastive loss. At the same time, the reduced attention matrix is employed to route features in both domains, so that source relations maintain in the synthesis. We validate our proposed method in three different I2I datasets, showing that it increases the image quality without adding learnable parameters. Codes are available at https://github.com/sapphire497/query-selected-attention.

1. Introduction

In image-to-image (I2I) translation, an input from the source domain X is mapped into the target domain Y while keeping its original content from unnecessary modifica-
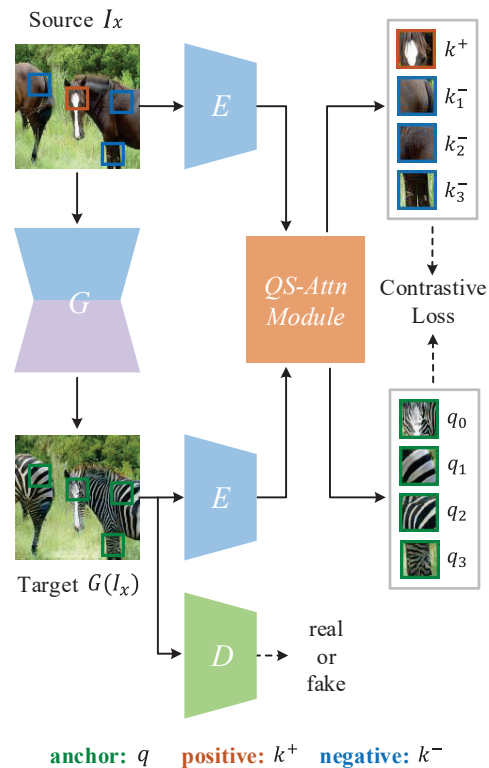
Figure 1. The overall structure of our model. The source domain image I_x is translated by the generator G into a target domain image $G(I_x)$. The encoder E extracts features from these two images, then the $QS\text{-}Attn$ module selects significant features to establish the contrastive loss. We also use a discriminator D to construct the adversarial loss.

tions. The translation is usually achieved by a generator G in the structure of auto-encoder with its output constraining by a discriminator D, so that it fulfills the requirement of the domain Y. In many I2I tasks, paired data are impossible to obtain, hence G can not be directly guided by the real image in Y. Ensuring that the output takes the input content is important for increasing its quality. Typical methods pro-
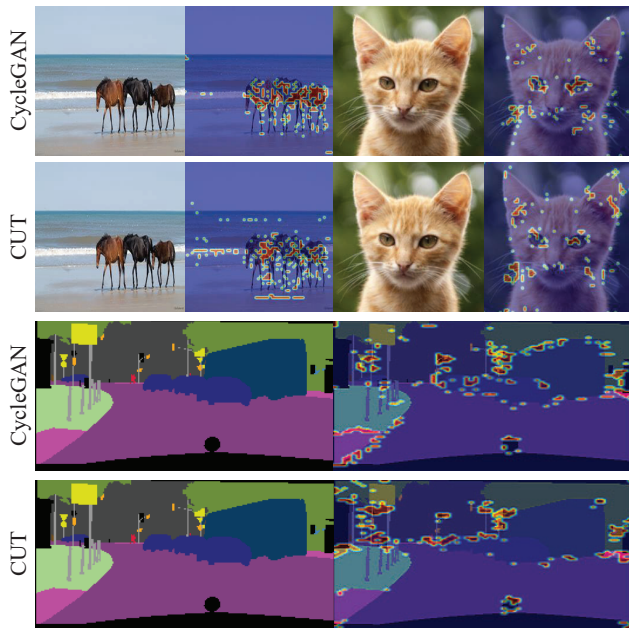
CycleGAN

CUT

CycleGAN

CUT

Figure 2. Visualization of feature significance metric from pretrained CycleGAN and CUT on three datasets. We show the entropy of attention matrix for each location, the warmer color indicates that the entropy is smaller. For each dataset, on the left column are the input images, and on the right column are the entropy visualizations of two models.

posed in [25, 42, 47] translate the result back into domain X by another generator G', and set up the cycle consistency penalty between the input and the final output. Although improving the quality, they introduce two generators and discriminators, which greatly increase the training costs.

Recently, CUT [34] incorporates the contrastive learning between the cross domain features from G's output and input. The key idea is to constrain the features from encoder E, requiring those from the same location to be close, compared with those from different locations. Removing the *QS-Attn module*, Fig. 1 illustrates the overall structure of CUT. An anchor point at a random position is selected from the features of the translated image, then one corresponding positive and many negatives are also sampled from the features of input. The contrastive loss is computed for the anchor so that the model maximizes the mutual information between the corresponding features. Note that CUT has only a single direction. Therefore, only one G is needed, and the training cost is reduced. The image quality is greatly improved, showing that contrastive loss across domains is useful in I2I.

However, there are still two issues ignored by CUT, which can be potentially improved. First, it does not select the anchors with purpose in the contrastive learning. Since each of them represents a small patch in the orig-

inal image resolution and many of them may not reflect any domain characteristics relevant for I2I. We argue that only those containing significant domain information need to be edited, and the contrastive loss imposed on them are more meaningful to guarantee the consistency across domains. Second, each anchor feature has only limited receptive field, and it does not consider its relation with other locations. This relation provides valuable cues to keep the source content stable and make the translation relevant.

We consider the above two issues in a simple way, inserting the *QS-Attn module* into the model as Fig. 1 shows, without bringing in extra model parameters. To evaluate the feature significance at different locations, we directly utilize features from E as both queries and keys to calculate the attention matrix in the source domain, then the distribution entropy is computed as a metric. Intuitive illustrations are provided in Fig. 2, in which such an entropy metric is visualized in the form of heat map. Particularly, given input images that need to be translated, we apply the encoder of pretrained CycleGAN [47] and CUT [34] models to obtain the features and calculate the attention matrix, and then the entropy is computed for each row of it. We sort the entropy in ascending order and show the smallest N points on the image. For the *Horse* and *Cat* images, the entropy values on the body of horse and the face of cat are smaller. For the *Label* image, the points mainly locate at the edges of categories. Consequently, the entropy can be a metric to measure how important the feature is in reflecting domain characteristics, hence we can impose the contrastive loss on it, ensuring the accurate translation on the domain-relevant features.

This paper intends to quantitatively measure the significance of each anchor feature and select the relevant ones for the contrastive loss according to the metric. Based on the previous analysis, we calculate the entropy of each row in the attention matrix and keep those with smaller entropy values. The remaining rows form the query-selected attention (QS-Attn) matrix, which consists of fewer queries, and they are further employed to route the value feature. Here the same matrix is multiplied with the values from both source and target domains, which implicitly keeps the feature relation in the source domain, avoiding excessive modifications on the result.

The contributions of this paper lie in following aspects:

- We propose a QS-Attn mechanism in I2I task. Our scheme is to choose the relevant anchor points, and use them as queries to attend and absorb features at other locations, forming better features suitable for contrastive learning. The QS-Attn keeps the simple design in CUT, and does not add any model parameters.

- We investigate different ways to quantify the signifi-

cance of the queries, to perform the attention, and to route value features in QS-Attn module, and find the entropy-based measurement and global attention for the cross domain value routing is the robust one.

- We do intensive experiments on the commonly used datasets, and show the proposed method achieves SOTA in most two domains I2I tasks.

2. Related Works

Image-to-image translation. GAN [2, 15, 24, 31] has the strong ability to describe the high dimensional distribution, therefore, has been widely used in the tasks of image synthesis like super-resolution [26], de-noising [4] and I2I. The I2I is first presented in Pix2Pix [21], and extended to high-resolution in Pix2PixHD [40], which can be regarded as a type of conditional GAN. The generator G consists of a pair of connected encoder-decoder, which translates an image from the source to the target domain. It is trained by paired data together with the adversarial loss from a target domain discriminator. However, unpaired I2I is more desirable, since the matched data across domains are impossible to be collected in most settings. CycleGAN [47] and DiscoGAN [25] achieve the I2I based on the unpaired data. They simultaneously train two different G, being responsible for the two directions of image translation. The cycle consistency is set up by successively using the two G for the opposite translation, and requiring the output to reconstruct the input source, which ensures G to employ the given content during translation, and maximizes the mutual information between the output and input source. The idea can be applied to multiple domains defined by several attributes [11], or utilized in the feature space [12, 19] in a flexible way. Meanwhile, many works [20,27,28] try to give the diverse translations by mixing the content and style from different images, or supporting the random sampling in the latent space. Together with the cycle consistency, translated images with the same content can appear different styles.

However, the cycle consistency is usually blamed for its strong constraint directly on the pixel [32, 44], which is not only unnecessary, but sometimes degrades the image quality. Except computing it in the feature level, another simple way is to perform the single direction translation. In this setting, the key issue becomes to keep the input content, and the extra loss term needs to be added. Distance-GAN [1] requires the generator to preserve the pixel level distance across two domains. GcGAN [13] links the source and target through a predefined geometry function. On the other hand, the feature level perceptual loss [22, 30] specified by a pretrained VGG [38] is widely adopted [5,14,30], which maintains the high-level semantic of the result. Nevertheless, features from a fixed layer of a pretrained model may not reflect the content which needs to be kept. At-

tnGAN [10] and GANimation [36] learn a foreground mask to guide the generator so that it realizes the translation in the relevant area. But they need extra parameters to estimate the foreground which definitely increase the model complexity. CUT [34] is the first attempt to incorporate the self-supervised contrastive loss into I2I, which significantly increases the translation quality. F-LSeSim [45] extends CUT by computing the self similarity within a local region, and imposes contrastive loss on it. However, it relies on features from VGG to measure the similarities, which reduces the training efficiency. We emphasize that both CUT and F-LSeSim do not intentionally select anchor features for contrastive loss, and their features still lacks large receptive field for representing image in source domain.

Self-supervised contrastive learning. Despite the great success of supervised learning, deep neural network is accused of its requirement on the large amount of the labeled training data. Recent studies of the self-supervised learning show its strong ability to represent an image without labels, particularly with the help of the contrastive loss [33, 41]. Its idea is to perform the instance level discrimination and learn the feature embedding, by pulling the features from the same image together and pushing those from different ones away. Recently, it has been investigated as a pre-training technique [3, 6, 8, 9, 16, 17], providing the initial model or the latent embedding for the down-stream task. The self-supervised learning has also been applied in image generation [7, 35]. SS-GAN [7] incorporates the rotation degree prediction as an auxiliary task for the discriminator, preventing the overfitting due to the limited data for real/fake binary classification. LT-GAN [35] trains an auxiliary classifier on top of the embedding in the discriminator, to classify whether the two pairs of fake images have the same perturbations on the sampling noise vectors. Besides CUT [34], the work [23] also adopts the self-contrastive learning for I2I. It employs non-local attention matrix to warp target image to the source pose, and requires feature from warped image to be close to the source through the contrastive loss. None of them chooses anchors, or utilizes the relation in the source domain like our method.

3. Methods

3.1. Preliminaries on CUT

In the task of I2I, given an image $I_x \in \mathbb{R}^{H \times W \times 3}$ from source domain X, the model aims to translate it into $G(I_x)$ in the target domain Y, having no obvious distinctions with the real image $I_y \in \mathbb{R}^{H \times W \times 3}$ in that domain. Generally, there are two auto-encoders, one $G_{X \to Y}$ for X to Y and the other $G_{Y \to X}$ for the reverse. CUT focuses on the single-direction translation from X to Y, so it only needs one generator G and one discriminator D, therefore, the subscript is omitted. The objective of adversarial loss L_{adv} can be
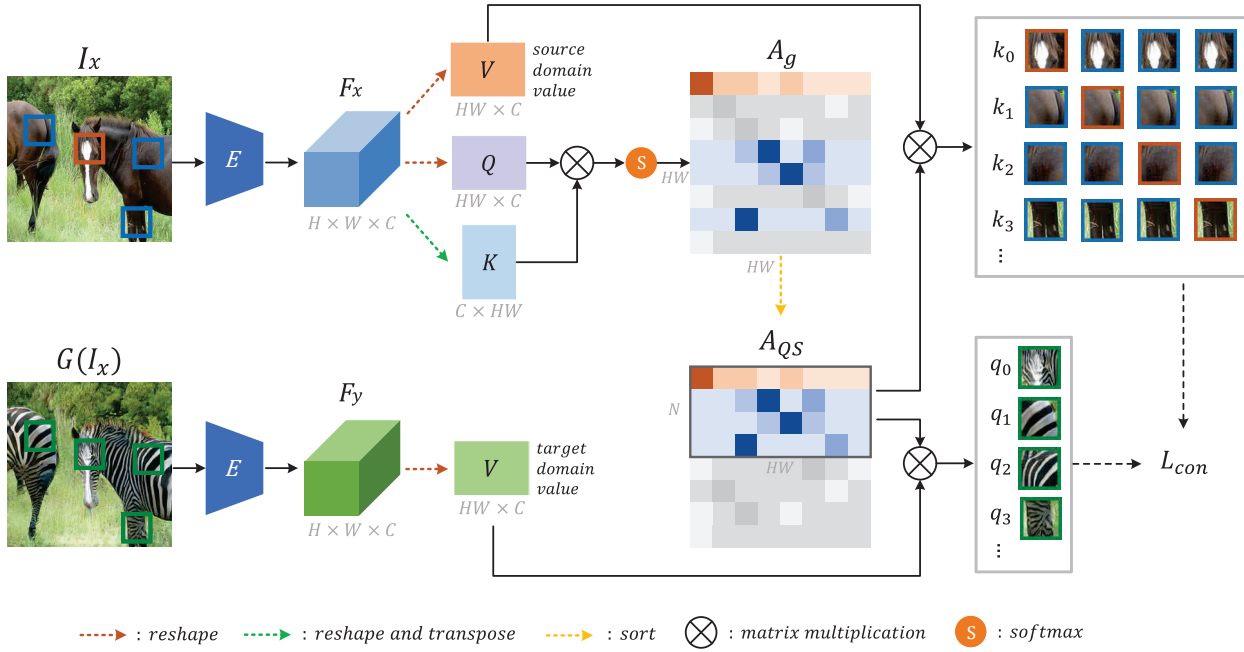
Figure 3. The details of QS-Attn. The encoder E extracts features F_x and F_y from I_x and $G(I_x)$, then F_x is reshaped and computed to derive the attention matrix A_g. Each row in A_g is sorted by its metric of the significance, and the selected N rows forming the A_{QS}. We further apply A_{QS} to route both source and target domain value features, and obtain positive, negative and anchor features to construct the contrastive loss L_{con}. Positive and negatives are from the real image I_x, while anchors are from the translated image $G(I_x)$. The patches in orange, blue and green indicate the positive, negative and anchor, respectively.

computed as follows.

$$L_{adv} = \mathbb{E}_{I_y \in Y} \log \mathrm{D}(I_y) + \mathbb{E}_{I_x \in X} \log(1 - \mathrm{D}(\mathrm{G}(I_x))) \tag{1}$$

Besides the L_{adv} in Eq. (1), CUT takes advantage of the first half of G as an encoder E to provide the extra constraint for the output from G. Basically, E compares the feature similarities across different domains. This efficient scheme is slightly different from previous work [48], in which another encoder E is employed, and is also adopted in [29]. It extracts features from both I_x and $G(I_x)$, and establishes the self-supervised contrastive loss in Eq. (2),

$$L_{con} = -\log\left[\frac{\exp(q \cdot k^+/\tau)}{\exp(q \cdot k^+/\tau) + \sum_{i=1}^{N-1} \exp(q \cdot k^-/\tau)}\right] \tag{2}$$

where q is the anchor feature from $G(I_x)$, k^+ is a single positive and k^- are $(N-1)$ negatives. Here τ indicates a temperature hyper-parameter. Note that the anchor q always locates in the fake image, and its positive k^+ is on the same location in the real image I_x. Besides, $(N-1)$ negatives k^- are randomly selected in I_x. The gradient of L_{con} only applies on the anchor q to train the parameters in G, while it is detached on k^+ and k^-, so that G is guided for the single direction of domain translation.

The full objective is expressed as follows.

$$L_G = L_{adv} + L_{con}^X + L_{con}^Y \tag{3}$$

where L_{con}^X is the contrastive loss defined in Eq. (2), and L_{con}^Y is the identity loss, in which the positive k^+ and negatives k^- are from a real target domain image I_y, and the anchor q is from $G(I_y)$. This identity loss guarantees the features from $G(I_y)$ are similar with features from I_y, preventing G from making changes on the target domain images.

3.2. QS-Attn for Contrastive Learning

As is shown in Fig. 1, we keep the simple setting like CUT. E is applied to extract features from I_x and $G(I_x)$. These features are supposed to establish the contrastive loss L_{con} defined in Eq. (2). The key module QS-Attn, setting up L_{con} across two domains, is illustrated in Fig. 3. Instead of the simple random strategy, we employ the idea of attention, which first compares a given query with keys, and then selects the query based on the comparison result. However, we do not use any separated projection head for query, key and value like the common attention, therefore, without adding extra model parameters in both G and D. Details for QS-Attn are given in the following two sub-sections.

3.2.1 Attention for query selection.

CUT randomly selects the anchor q, positive k^+ and negatives k^- to compute the contrastive loss in Eq. (2), which is potentially inefficient, because their corresponding patches may not come from the domain-relevant region, *e.g.* the horse body in the *Horse → Zebra* task. Note that some features do not reflect the domain characteristics, they tend to be kept during the translation. Therefore, the L_{con} imposed on them is not vital for G. Our intention is to choose the anchor q, and compute L_{con} on the significant ones which contains more domain-specific information.

Global attention. Based on the above observation, we aim to define a quantitative value for each potential location, which reflects the significance of the feature. The quadratic attention matrix is adopted, since it exhaustively compares each feature with all other locations, it accurately reflects the similarities with others, as is shown in Fig. 3. Particularly, given a feature $F_x \in \mathbb{R}^{H \times W \times C}$ in the source domain, we first reshape it into a 2D matrix $Q \in \mathbb{R}^{HW \times C}$, and then multiply it by its transposed $K \in \mathbb{R}^{C \times HW}$. Then we give each row of the matrix to the softmax function, leading to a global attention matrix $A_g \in \mathbb{R}^{HW \times HW}$. Consequently, significant features can be measured according to the entropy H_g of each row in A_g, which is computed as in Eq. (4).

$$H_g(i) = -\sum_{j=1}^{HW} A_g(i,j) \log A_g(i,j) \quad (4)$$

Here i and j are the indexes of the query and key, corresponding to the row and column in A_g. When $H_g(i)$ approaches to 0, it means that in the i-th row, only a very few key locations are similar with the i-th query. Hence we assume that it is distinct enough and is important to be constrained by L_{con}. To select all the significant queries, the rows of A_g are sorted by the entropy H_g in the ascending order, and the smallest N rows are selected as the QS-Attn matrix $A_{QS} \in \mathbb{R}^{N \times HW}$. Note that A_{QS} is fully determined by the features in I_x, and has no relation with $G(I_x)$.

Local attention. Though non-local attention can obtain the global context, it smooths out the detailed context surrounding the queries. Local attention measures the similarity between a query and its neighboring keys within a constant window of $w \times w$ and stride of 1, which can capture the spatial interactions in local regions, and reduce the computation cost. Given a reshaped query matrix $Q_l \in \mathbb{R}^{HW \times C}$, we multiply it by local key matrix $K_l \in \mathbb{R}^{HW \times w^2 \times C}$ and send it to the softmax function, leading to a local attention matrix $A_l \in \mathbb{R}^{HW \times w^2}$. The local entropy H_l is computed in each row as in Eq. (5).

$$H_l(i) = -\sum_{j=1}^{w^2} A_l(i,j) \log A_l(i,j) \quad (5)$$

Here i and j are the indexes of the query and key. We select the smallest N rows in A_l by sorting H_l in the ascending order to form A_{QS}. For the value routing, We also locate the N indexes in local value matrix $V_l \in \mathbb{R}^{HW \times w^2 \times C}$ and get the selected value matrix $V_{ls} \in \mathbb{R}^{N \times w^2 \times C}$.

3.2.2 Cross domain value routing for contrastive learning.

The reduced A_{QS} is used as the attention matrix to route the value features from both source and target domains. Here we emphasize that A_{QS} captures the global or local relation by comparing the query with keys, and it provides useful high-order descriptions about the shape and texture of I_x. Using it to route features helps to enlarge the receptive field of the selected queries, so that better features, which consider the context of I_x, can be formulated. Moreover, the relation defined by A_{QS} is also required to be kept during the image translation. So A_{QS} is imposed on the features from both I_x and $G(I_x)$, routing the corresponding value to form the anchor, positive and negatives. One positive and $(N-1)$ negative features are located in the real image I_x. N anchors are from the fake image $G(I_x)$. We establish the self-supervised contrastive loss as Eq. (2), using these features to constrain the translation.

4. Experiments

4.1. Implementation Details

Datasets. Our model is trained and evaluated on *Cityscapes*, *Horse → Zebra*, and *Cat → Dog* datasets. *Cityscapes* contains street scenes from German cities, with 2,975 training images. *Cat → Dog* are from AFHQ Dataset [12], which consists of 5,153 and 4,739 training images for cat and dog, respectively. The images for *Horse → Zebra* are provided in [47], which contains 1,067 and 1,334 training images for horse and zebra, respectively. For all experiments, the resolution of input and generated images is 256×256. The initial resolution of images in *Cityscapes* and *Cat → Dog* dataset are 2048×1024 and 512×512, and we resize them to 256×256 in our experiments.

Training details. We build our model with a ResNet-based generator and a PatchGAN discriminator [39], and compare it with CUT in the same setting on three aforementioned datasets. The number of rows in attention matrix is set to 256, and the dimension of anchor, query, key features for computing contrastive loss is 256. We adopt the multi-layer feature extraction in CUT, which takes the features from five layers. Considering the high cost in computing the global attention matrix, we propose to apply the QS-Attn on the last two layers' feature in the encoder, but we still make an extra comparison with the model in which all layers are applied it. Details can be found in the supple-
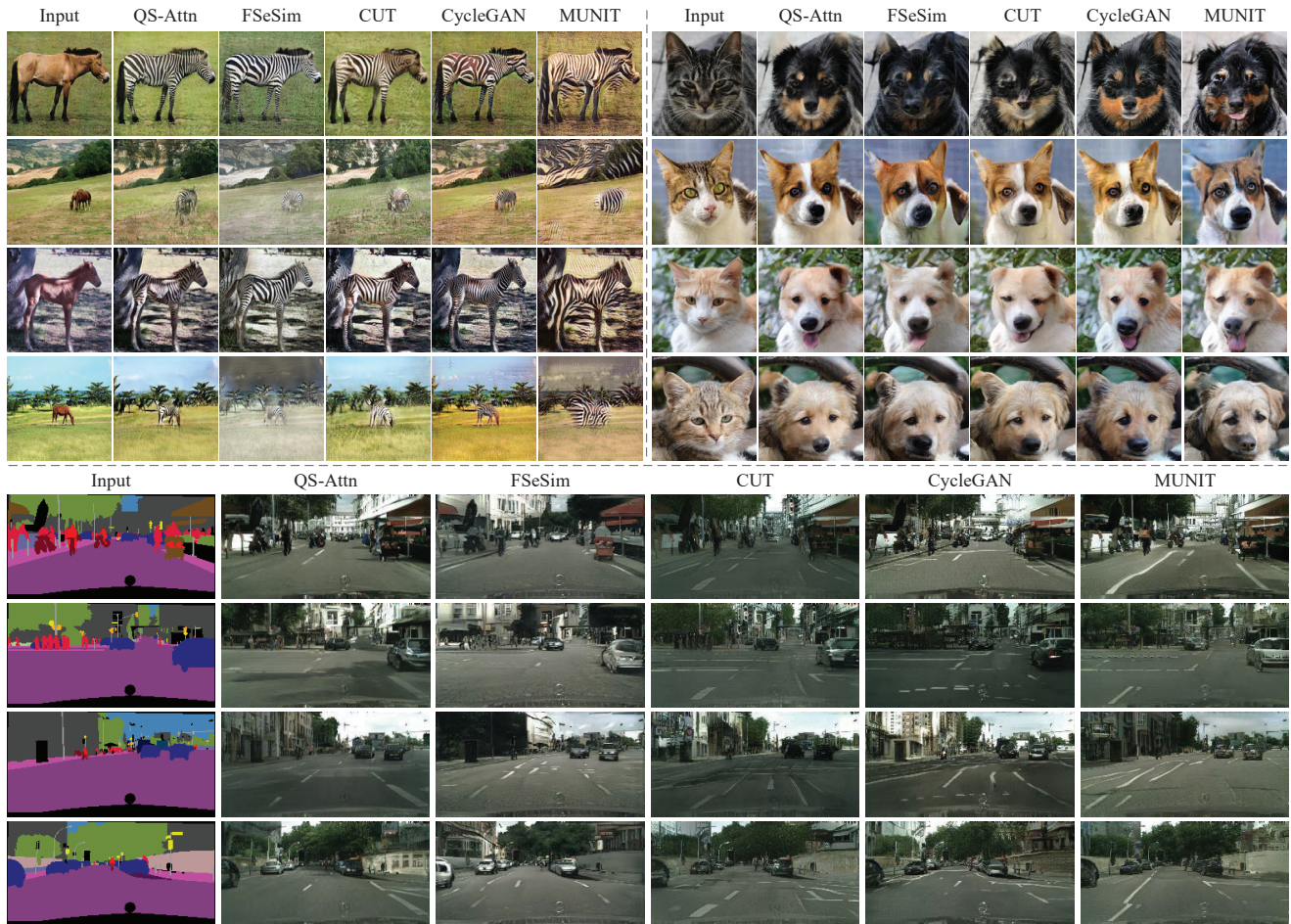
Figure 4. Visual results comparison with other methods. We compare our model with FSeSim, CUT, CycleGAN and MUNIT on three benchmark datasets. In the results of QS-Attn, the translated images of *Horse → Zebra* and *Cat → Dog* are from QS-Attn(Global), and the results of *Cityscapes* are generated using QS-Attn(Global+Local). More results can be found in the supplementary materials.

Method	Cityscapes				Cat→Dog		Horse→Zebra	
	mAP↑	pixAcc↑	classAcc↑	FID↓	SWD↓	FID↓	SWD↓	FID↓
CycleGAN	20.4	55.9	25.4	76.3	19.5	85.9	39.1	77.2
MUNIT	16.9	56.5	22.5	91.4	24.4	104.4	50.7	133.8
CUT	24.7	68.8	30.7	56.4	12.9	76.2	31.5	45.5
FSeSim	22.1	69.4	27.8	54.3	13.8	87.8	37.2	43.4
QS-Attn(Global)	25.5	79.9	31.2	53.5	**12.8**	**72.8**	**30.3**	41.1
QS-Attn(Local)	26.2	80.5	31.9	**48.8**	13.3	79.3	31.2	**38.6**
QS-Attn(Local+Global)	**27.9**	**81.4**	**32.6**	50.2	13.2	80.0	31.9	42.3

Table 1. Quantitative comparison with other methods. The last three rows are our models with different settings, details are illustrated in Sec. 4.2. The best performance is indicated in **bold**.

mentary materials.

Evaluation metrics. We use Fréchet Inception Distance (FID) [18] and Sliced Wasserstein Distance (SWD) [37] to evaluate the quality of translated images. FID and SWD both measure the distance between two distributions of real and generated images, and lower indicate the generated image is similar to the real one. For *Cityscapes* dataset, following [34], we apply semantic segmentation to the generated images using DRN [43], and compute mean average precision (mAP), pixel-wise accuracy (pixAcc), and aver-
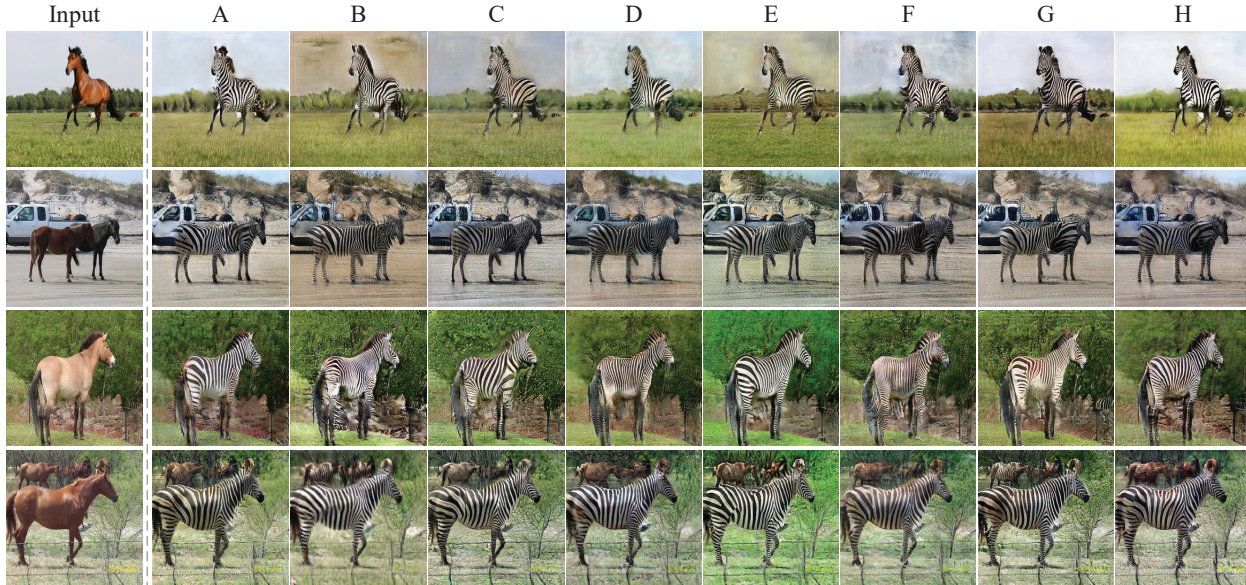
Figure 5. Qualitative ablation results. The leftmost column are input images, the remaining columns are translated images using model A-H. Details are illustrated in Tab. 3 and Sec. 4.3.

age class accuracy (classAcc), showing the semantic interpretability of the generated images. We calculate the metrics in the whole test set, in which *Cityscapes* contains 500 label images, *Cat* → *Dog* includes 500 cat images, and *Horse* → *Zebra* comprises 120 horse images.

4.2. Results

Quantitative and qualitative results. Tab. 1 compares our model with FSeSim [45], CUT, CycleGAN [47] and MUNIT [20]. There are three settings in our model, *Global*, *Local* and *Local+Global*. *Global* means sorting the global attention matrix A_g by entropy to form the QS-Attn matrix A_{QS}, and *Local* means sorting the local attention matrix A_l to form A_{QS}. Furthermore, in order to utilize both local and global context, *Local+Global* applies A_l for query selection and A_g in value routing. The rows of A_g are sorted by the local entropy H^l in the ascending order, and the smallest N rows are selected as A_{QS}. The ablation study of the three models is illustrated in Sec. 4.3 and the code is provided in the supplementary materials. For the metric of FID, the translated results of our model are more realistic than other methods on three datasets. Our model also performs better on mAP, pixAcc and classAcc on *Cityscapes* dataset. Moreover, our method **does not** add extra model parameters in both G and D, and uses the same architecture of G as CycleGAN.

Visual results are shown in Figure 4. Compared to other methods, our QS-Attn model has the ability to translate the domain-relevant features accurately. Besides, QS-Attn achieves the background consistency in the tasks of *Horse* → *Zebra* and *Cat* → *Dog*.

	QS-Attn	FSeSim	CUT	CycleGAN
Q(%)	**45.0**	28.3	13.3	13.3
T(%)	**40.0**	30.0	11.7	18.3
C(%)	**58.3**	11.7	8.3	21.7

Table 2. User study statistics. The methods are compared in three aspects: image quality (Q), target domain conformity (T) and domain-irrelevant consistency (C).

User Study. To further evaluate the quality of translated images, we conduct a user study under human perception. so a user study reflecting human perception is important for evaluating visual quality. We compare our *Global* model with CUT, FSeSim, and CycleGAN on *Horse* → *Zebra* dataset. 60 participants are asked to compare the methods in three aspects: image quality (Q), target domain conformity (T) and domain-irrelevant consistency (C). Q means the reality and perception of images. T indicates whether the translated images have the features of the target domain. C refers to the domain-irrelevant pixels should remains the same compared to the source images, *e.g.*, the background in the *Horse* image. Statistical results are shown in Tab. 2.

4.3. Ablation Study

Attention and selection. In QS-Attn module, we apply attention, query selection and cross domain value routing. To evaluate their effects separately, we conduct the ablation study on *Horse* → *Zebra* dataset. Metrics are listed in Tab. 3 and qualitative results are shown in Fig. 5. **A** is our complete global model, including aforementioned three operations. In model **B**, queries are selected randomly and the

Method	Attention		Selection		Cross Domain	FID ↓	SWD ↓
	global	local	global	local			
A	✓		✓		✓	41.1	**30.3**
B	✓		×		✓	52.1	34.4
C	×		✓		×	61.1	37.7
D	✓		✓		×	43.3	32.1
E	✓		×		×	53.4	34.9
F: info	✓		✓		✓	43.9	33.3
G		✓		✓	✓	**38.6**	31.2
H	✓		✓		✓	42.3	31.9

Table 3. Quantitative results for ablation study. In configuration, $Attention$ means computing the QS-Attn matrix A_{QS} and adopting it to route the value feature; $Selection$ denotes selecting queries by entropy in attention matrix; $Cross\ Domain$ indicates that A_{QS} in the source domain is applied to route the value features from both source and target domains; $global$ and $local$ refer to using A_g and A_l, respectively. Model **A**, **G** and **H** are the 3 settings corresponding to the last 3 rows in Table 1.

corresponding QS-Attn matrix A_{QS} is applied to route the source and target values. Model **C** only computes the global attention matrix A_g to select queries, while it does not further use A_{QS} to route values. Model **A** outperforms model **B**, reflecting the effectiveness of entropy-sorted selection. The metric of **C** is worse than model **A** and **B**, indicating that only when the selected A_{QS} routes the values to establish the contrastive loss L_{con}, then the encoder learns to extract the significant features from images.

Self domain value routing. Model **D** and **E** in Tab. 3 route value features in self domain, *i.e.* there are two global attention matrices A_g^x and A_g^y, from source domain and target domain, respectively. Then, after the query selection, the source domain QS-Attn matrix A_{QS}^x and the target domain QS-Attn matrix A_{QS}^y route the value features from their own domain. In model **D**, queries are selected by sorting the entropy of A_g^x to form A_{QS}^x, and A_{QS}^y is composed of the selected queries in A_g^y, which has the same row index as A_g^x. Different from **D**, model **E** selects queries randomly. The two models both route the values separately by the corresponding A_{QS}^x and A_{QS}^y. Model **D** is not good as **A**. It demonstrates that compared with self domain routing, cross domain routing can establish closer correlation between source and target domain, for A_{QS}^x captures the global relation of I_x and imposes it on the features from $G(I_x)$.

Informer. We also investigate other query selection strategies. Recently, Informer [46] proposes an efficient self-attention mechanism. It introduces a max-mean measurement for each query, which is expressed as:

$$M(q_i; K) = \max_j \left(\frac{q_i k_j^T}{\sqrt{C}} \right) - \frac{1}{HW} \sum_{j=1}^{HW} \left(\frac{q_i k_j^T}{\sqrt{C}} \right) \quad (6)$$

where $q_i \in \mathbb{R}^C$ is the i-th query in the matrix $Q \in \mathbb{R}^{HW \times C}$, $k_j \in \mathbb{R}^C$ is the j-th key in the matrix $K \in \mathbb{R}^{C \times HW}$. Each query measures the similarity with all keys, and obtains a score $M(q_i; K)$. Then Top-N queries are selected by sorting $M(q_i; K)$ in descending order, which are relatively distinct. Model **F** in Tab. 3 selects Top-N queries adopting the above approach, and uses the corresponding A_{QS} to route both source and target domain value features. It achieves good results and metric, showing that the max-mean measurement is also effective for query selection.

Local and global attention. Model **G** and **H** applies local attention for A_l, and computes H_l for query selection. Differently, **G** employs selected rows in A_l to form A_{QS} in value routing, but **H** uses global attention matrix A_g. Although **G** achieves the best performance in FID, the visual quality is worse than **H** and **A**, suggesting that global value routing helps to reconstruct the texture of images.

5. Conclusion

This paper proposes a QS-Attn module for cross domain contrastive learning in the task of I2I. Instead of randomly selecting the anchor, positive and negatives to compute the contrastive loss, we measure the significance of source domain features and select them based on a metric so that the constraint becomes more relevant for the domain translation. We first compute an attention matrix using the features from real images in the source domain, and then measure the entropy of every query in it. Those with smaller values are considered to be distinct, therefore being selected. The remaining significant queries are kept, resulting in a row reduced attention matrix, which is further employed to route features in both the source and target domains. The cross domain routing strategy not only enlarges the receptive field of the selected features, but also helps the output to retain the relation in the input image. We show the effectiveness of QS-Attn module on popular domain translation datasets and perform intensive ablation studies.

References

[1] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. *arXiv preprint arXiv:1706.00826*, 2017. 3

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 3

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 3

[4] Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3155–3164, 2018. 3

[5] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017. 3

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[7] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12154–12163, 2019. 3

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020. 3

[10] Xinyuan Chen, Chang Xu, Xiaokang Yang, and Dacheng Tao. Attention-gan for object transfiguration in wild images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 164–180, 2018. 3

[11] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 3

[12] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3, 5

[13] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2427–2436, 2019. 3

[14] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 3

[15] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 3

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 3

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 6626–6637. Curran Associates, Inc., 2017. 6

[19] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018. 3

[20] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 3, 7

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 3

[23] Taewon Kang, Soohyun Kim, Sunwoo Kim, and Seungryong Kim. Online exemplar fine-tuning for image-to-image translation. *arXiv preprint arXiv:2011.09330*, 2020. 3

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 3

[25] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pages 1857–1865. PMLR, 2017. 2, 3

[26] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-

realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 3

[27] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 3

[28] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *arXiv preprint arXiv:1703.00848*, 2017. 3

[29] Rui Liu, Yixiao Ge, Ching Lam Choi, Xiaogang Wang, and Hongsheng Li. Divco: Diverse conditional image synthesis via contrastive generative adversarial network, 2021. 4

[30] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 768–783, 2018. 3

[31] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[32] Ori Nizan and Ayellet Tal. Breaking the cycle-colleagues are all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7860–7869, 2020. 3

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[34] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020. 2, 3, 6

[35] Parth Patel, Nupur Kumari, Mayank Singh, and Balaji Krishnamurthy. Lt-gan: Self-supervised gan with latent transformation detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3189–3198, 2021. 3

[36] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833, 2018. 3

[37] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011. 6

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[39] Jaemin Son, Sang Jun Park, and Kyu-Hwan Jung. Retinal vessel segmentation in fundoscopic images with generative adversarial networks. *arXiv preprint arXiv:1706.09318*, 2017. 5

[40] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 3

[41] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 3

[42] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017. 2

[43] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017. 6

[44] Yihao Zhao, Ruihai Wu, and Hao Dong. Unpaired image-to-image translation using adversarial consistency loss. In *European Conference on Computer Vision*, pages 800–815. Springer, 2020. 3

[45] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. The spatially-correlative loss for various image translation tasks. In *CVPR*, 2021. 3, 7

[46] Hao-Yi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *ArXiv*, abs/2012.07436, 2020. 8

[47] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2, 3, 5, 7

[48] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. 11 2017. 4