Causal-SETR: A SEgmentation TRansformer Variant Based on Causal Intervention [*]

Wei Li and Zhixin Li(✉)

Guangxi Key Lab of Multi-source Information Mining and Security,
Guangxi Normal University, Guilin 541004, China
lizx@gxnu.edu.cn

Abstract. We present a novel SEgmentaion TRansformer variant based on causal intervention. It serves as an improved vision encoder for semantic segmentation. Many studies have proved that vision transformers (ViT) can achieve a competitive benchmark on these downstream tasks, which shows that they can learn feature representations well. In other words, it is good at observing the instance from the image. However, in the human visual system, to recognize the objects in the scene, it is necessary to observe the objects themselves and introduce some prior knowledge for producing higher confidence results. Inspired by this, we introduced a structural causal model (SCM) to model images, category labels, and context. Beyond observing, we propose a causal intervention method by removing the confounding bias of global context and plugging it in the ViT encoder. Unlike other sequence-to-sequence prediction tasks, we use causal intervention instead of likelihood. Besides, the proxy training objective of the framework is to predict the contextual objects of a region. Finally, we combine this encoder with the segmentation decoder. Experiments show that our proposed method is flexible and effective.

Keywords: Causal intervention · Vision transformer · Semantic segmentation.

1 Introduction

Semantic segmentation divides visual input into different semantically interpretable categories, which is a challenging task requiring accurate prediction of the object category, shape, and location. Both convolutional-based encoders [19,5,25] and transformer-based encoders [10,38,17] are good at telling us "**what**", but not "**why**". In particular, once the input image has been fed to the encoder, the rich and effective feature representation can be learned to provide a high confidence probability $P(Y)$. Furthermore, some empirical investigations

[*] This work is supported by National Natural Science Foundation of China (Nos. 62276073, 61966004), Guangxi Natural Science Foundation (No. 2019GXNSF DA245018), Guangxi "Bagui Scholar" Teams for Innovation and Research Project, and Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing.

[36,11,26,37] use prior knowledge, graph neural networks and other techniques to learn the co-occurrence probability $P(Y|X)$ between objects and integrate it with the probability learned by the encoder for joint learning in order to increase the prediction probability. However, Wang et al. [30] raise questions about the validity of the co-occurrence probability learned by machine, and they think the machine usually fails to describe the exact visual relationships, or, even if the prediction is correct, the underlying visual attention is illogical. Improving the capabilities of the segmentation systems by acquiring higher co-occurrence information with a better degree of confidence is thus a crucial issue.Contextual information is crucial for semantic segmentation tasks. Some methods based on graph convolutional networks (GCN) learn the rich context to improve the feature representation capability. Moreover, Zheng et al.[38] provide a rethinking of the segmentation model and contribute a new encoder-decoder architecture built by pure transformers. This architecture does not involve spatial resolution down-sampling, but rather global context modeling at each layer of the encoder transformer. With the global context, they propose a new perspective to the semantic segmentation. As previously stated, the machine is incapable of describing the precise visual relationship. (For example, the "visual" simply conveys the "what" or "where" of a "person" or "car".) It is just a more descriptive symbol than its corresponding English word. When there is a bias, such as when more "car" areas than “human" regions co-occur with the term "road," visual attention is more likely to focus on the “car" region. These works [7,6,31] attempt to introduce unsupervised external features to obtain more robust co-occurrence relations, thus improving the segmentation performance. Contrary to human's recognition system, current deep learning approaches cannot yet extract or explain causality.

According to these causal theories [22,21,23], we intend to reconsider the segmentation based on the vision transformer model design and contribute a causal intervention attempt. In particular, we model images, category labels, and contextual information using SCM and eliminate confounder bias through causal intervention. Thus, we obtain contextual information regarding causality to drive the learning of more robust semantic features and the exploration of the deeper relationships between various objects. In addition, we design a fusion module for integrating the original feature and the causality context, so intervening in the learnt features and making the learning process more like the human learning process. It's worth noticing that the proposed module is plug and play. We can easily plug it into other downstream tasks. In summary, we make the following contributions in this paper:

- We introduced the structural casual model to model images, category labels, contextural information, and removed the observation bias by causal intervention. Thus, we get the contextural causality information, which collects more robust semantic relations.
- We incorporated external knowledge into the processing of causal intervention as well as further guided the ViT model to provide a more robust feature representation.

– We designed a fusion module for integrating the original feature and the causality context, which performs reasoning and directs the downstream task (*e.g.,* semantic segmentation) to explore more causality interconnections.

To demonstrate the efficacy of the proposed method, we duplicated these approaches into an end-to-end training network and performed extensive experiments on the benchmark of semantic segmentation. Experiments demonstrate that our proposed methods are practicable and efficient.

2 Related work

2.1 Vision Transformer

The most related to our work is the vision transformer (ViT) [10] and its variants [28,32]. ViT treats an image as a set of fixed size (i.e., 16 × 16) and non-overlapping patches, then directly feeds them to a transformer architecture. Thus, it converts the dense prediction task to a sequence-to-sequence task. Compared to CNNs, it achieves a competitive speed-accuracy tradeoff on classification. However, ViT requires large-scale training datasets (i.e., JFT-300M). DeiT [29] adapts the knowledge distillation for reducing the complexity and finetuning the ViT, allowing ViT to be effective using the smaller ImageNet-1K dataset. We noticed that ViT lefts the results of image classification. However, it is still unsuitable for use as a general-purpose backbone on dense prediction tasks or handling high-resolution images due to its low-resolution feature map and the quadratic increase in complexity with image resolution. DETR [3], SETR [38] directly upsampling or deconvolution the features but with dissatisfied performance in detection and segmentation respectively. As far as we know, no one has tried to introduce the perspective of causality into ViT for semantic segmentation. Empirically, our proposed approaches are effective and flexible, achieving a new state-of-the-art in semantic segmentation task.

2.2 Causality in vision

Due to the fact that deep learning is an effective yet unexplained black box, more and more academics are attempting to combine its complementary strengths. Causal inference [22,21,23] has been researched in several domains, including classification [4,20], adversarial learning [14,15], and reinforcement learning [9,2]. The most related to our work is VC R-CNN [30]. They constructed a causal region of interest (RoI) using Faster R-CNN [27] and then use this contextual RoI further to improve the performance of several multimodal downstream tasks, including image caption (IC), visual question answering (VQA), vision commonsense reasoning (VCR).

The core idea between ours and VC R-CNN is backdoor adjustment solution. However, they did not report the potential interest in semantic segmentationfield. We observed that semantic segmentation tasks also require causal contextual information for advancement. Due to the task gap, we cannot directly introduce the causal RoI information [30] as context in our investigation. Therefore, we reconstructed a structural causal model on the benchmark of semantic segmentation tasks. Despite the fact that we both intend to mine the rich contextual via a backdoor modification approach, VC R-CNN uses a backdoor adjustment method to eliminate the visual bias caused by the model's "observing" behavior. Consequently, VC R-CNN may learn “common sense" without any external monitoring. It is important to note that the most significant distinction between VC R-CNN and our approach is that we additionally intervene on external knowledge to improve the performance of semantic segmentation.

CONTA [35] proposes a contextual adjustment network to improve the semi-supervised semantic segmentation benchmark. Similar to the case with the causal RoI features [30] , we cannot directly utilize the CONTA-supplied SCM. On the one hand, the backbone we use is a ViT rather than a CNN, and on the other hand, the general paradigm for weakly supervised semantic segmentation tasks does not correspond to the paradigm used for fully supervised image recognition tasks. We aim to improve the performance of ViT via adapting the backdoor adjustment solution for semantic segmentation tasks.

Hybridization effect will result in harmful bias, mislead attention module to learn false correlation in data, and consequently reduce the model's generalizability. However, Xu et al. [34] think that confounding is unobservable. Thus, they propose a novel attention mechanism: causal attention (CATT) which can eliminate the confounding effect in existing attention-based vision-language models. Unlike CATT, we employ backdoor adjustment solution as opposed to front-door adjustment solution. Furthermore, we focus on designing a generic ViT architecture via causal intervention, not a attention mechanism.

In summary, we aim to model the semantic segmentation tasks in detail using backdoor adjustment. VC R-CNN reported the inspiring performance in IC, VQA and VCR tasks. Similarly, CONTA reported the good performance in semi-supervised semantic segmentation task. They eliminate the observation biases from within the model using backdoor adjustment. However, we extracted some common sense from an external knowledge dataset which are presented as textual data, and introduced them into the backdoor adjustment processing. For segmentation tasks, contextual information plays an important role, thus, we convert the external knowledge (textual data) to visual commonsense features using GCN. Afterthat, we notice that ViT is a unified framework for modeling language and vision, and we make an attempt to rethink the advantage of causal intervention in the ViT-based model. Different from CNN, ViT lacks some inductive bias (*e.g.,* invariance, local connectivity, weight sharing) due to the framework design. Therefore, we complement external knowledge with the strengths of ViT from a causal perspective. As a result, we further improve the benchmark in segmentation tasks.

3 Methods

We attempt to intervene in the feature representation learned by the vision transformer encoder, thus, obtaining more explanatory contextual information, which improves the performance of semantic segmentation tasks. For example, the deep encoder learns contextual information with observation bias from the dataset (*e.g.,* if there are “car", “road" and “person" co-occur in an input image, the encoder is more likely to focus on the common co-occurrence relationships in the dataset.). Perhaps the classification result is correct, but the underlying visual attention is not reasonable. To address this, we propose the causal-intervention-based framework for obtaining a more causal context. The overview of the framework for semantic segmentation is shown in Fig. 1. It is worth noting that our proposed method can be used plug and play on any transformer-based encoders and is compatible with downstream recognition tasks.

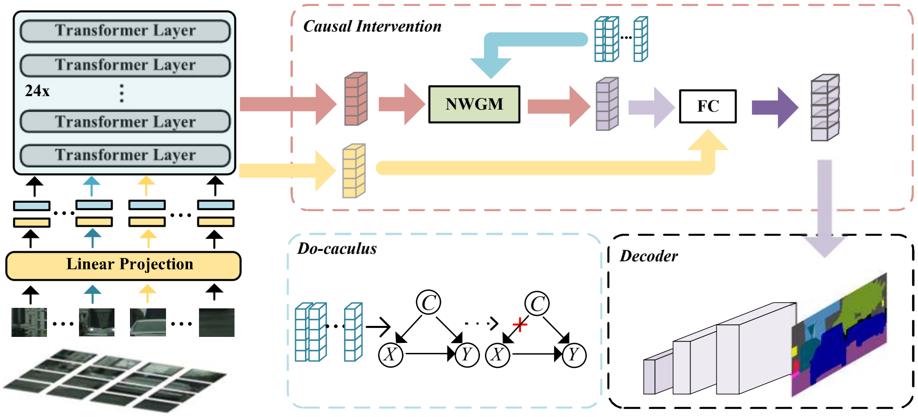

Fig. 1: Overview of our proposed semantic segmentation framework.

Visual attention is effective at learning the correlation $(P(Y|X))$ between objects. However, it is limited to the fixed input image. Therefore, it can only learn the explicit correlation in this image. In other words, visual attention is incapable of observing nonexistent objects in the input image. It disregards the implicit causality that causes the observation bias to confound the existence of objects X and Y.

To mine the implicit causality, we first build the confounder set $C \in \mathbb{R}^{N \times D}$. N represents the number of objects in the datasets,, and D is the dimension of the middle output produced by the feature encoder. Besides, we build another $C^{'}$ using external knowledge. Then, we “take” the objects C from other context, and “put” them around X and Y for testing if X causes the existence of Y when given C. The operators (“take” and “put”) are the paradigm of intervention, implying that the probability of C depends on human intervention, but is independent on X or Y. By intervening, we force the conversion of the correlation observed only in the fixed image to a global or external causal-based context. $(P(Y|X) \rightarrow P(Y|do(X)))$

More intuitively, human will not make corresponding inferences just based on what they “see” in front of eyes. That is what we are different from the machine’s visual recognition system. For example, we always keep rethinking or imaging “If there are other objects C , will object X still causes object Y?” instead of the passive observation: "If there is object X, how likely there will exist object Y? ". Thanks to intervention, we convert $P(Y|X)$ to $P(Y|do(X))$. We simulate the $do - calculus$ by "taking" non-local context that even might not be in the input image, "putting" them around pairs of objects that we want to intervene.

3.1 Structural Causal Model

As shown in Fig. 2, we intuitively demonstrate the principle of *do-calculus*. Specifically, we formulate causalities among observed objects X, confounder set C, and observed objects Y with a structural causal model. The symbol “→" denotes the causalities between two nodes (*e.g., X* causes Y).

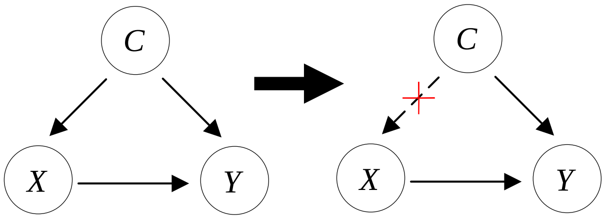

Fig. 2: Modeling the causalities by SCM

$\boldsymbol{C} \rightarrow \boldsymbol{X}$. It is widely known that context C affects the performance of the semantic segmentation model. In other words, C guide the model to what or where is “car", “road", and “person" in an image. We can hardly ever build a generative context for $C \rightarrow X$, but we will introduce an ingenious method to extract the links suitable for segmentation tasks

$\boldsymbol{X} \rightarrow \boldsymbol{Y}$. The link between X and Y denotes X causes Y (*e.g.* , if X exists in the image, what is the probability of Y exists in the same image?). It is learned by the likelihood: $P(Y|X)$. Although we have seen many successful convolutional-based methods make great progress by alleviating this likelihood, we still firmly believe that this is biased.

$\boldsymbol{X} \leftarrow \boldsymbol{C} \rightarrow \boldsymbol{Y}$. The confounder set C can be interpreted as a generic contextual corpus. It will cause both of X and Y by “taking” the implicit context that can not be observed in the local images and “putting” it to the local receptive field. However, it might leads to spurious correlations by only learning from the likelihood which is formulated as:

$$P(Y|X) = \sum_{c} P(Y|X, c)P(c|X) \tag{1}$$

Where the context C introduces the observational bias $P(c|X)$. We are more concerned with the prediction of Y, so the causal link between C and X is not the major link we need to focus on. If we intervene X, the causal link between C and X is cutoff. We apply the Bayes rules again. Thus, we have:

$$P(Y|X) = \sum_{c} P(Y|X, c)P(c) \tag{2}$$

Where $P(c)$ denotes the probability of prior label, it deliberately forces X to incorporate every c fairly. $c \in C$ is the set of the objects from the contextual corpus.

3.2 Deconfounding Bias

Where will we intervene in the human reasoning system when we find that confounding factors interfere with the output results? The optimal solution defined by mathematics is to average the different confounders and give the maximum weight to the more reliable and fewer error signals [1].

Therefore, we approximate the confounder set $C_{\text{internal}} = \{c_1, c_2, \ldots, c_n\}$, where c_n is the $N \times d$ matrix, N denotes the category size in the datasets (*e.g.,* $N = 19$ in the cityscapes dataset) , and d is the averaged mask of the i-th category features produced by the ViT encoder. In another word, C_{internal} is produced by the model itself which means to deconfound the internal bias.

Furthermore, we acquire an **external knowledge** set $\mathcal{E}$ from Visual Genome dataset [16]. However, the Visual Genome dataset consists of 30K object categories for the specific downstream task. For the semantic segmentation task, we attempt to mine the co-occurrence relationships about different object categories that appear in Cityscapes and ADE20K. Specifically, we get the subset $\mathcal{E}^{\text{external}} \in \mathbb{R}^{C \times C}$ of $\mathcal{E}$. C is set to 150 (ADE20K contains 150 object categories and overlaps with Citiscapes). $\mathcal{E}^{\text{external}} \in \mathbb{R}^{150 \times 150}$ is a 150×150 symmetric matric which means the relation pairs are symmetric. After that, we normalized the matrix elements to obtain D. $D_{ii} = \sum_{j=1}^{C} \mathcal{E}_{ij}$ The final $\mathcal{E}^{\text{external}} \in \{\mathcal{E}_{00}, \cdots \mathcal{E}_{ij}\}$ is calculated by $\mathcal{E}_{ij} = \frac{\mathcal{E}_{ij}}{\sqrt{D_{ij} D_{jj}}}$. To maintain the consistency between semantic relations and visual features, we introduce a graph structure $\mathcal{G} = (\mathcal{N}^L, \mathcal{E}^{\text{external}})$, where $\mathcal{N}^L$ is produced by global vectors for word representation such as GloVe [24]. Besides, we feed it to two GCN layers to capture external semantic knowledge. Each GCN layer is fomulated by

$$H^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}} \mathcal{E}^{\text{external}} \widetilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}) \tag{3}$$

Where $\widetilde{D}$ denotes the degree of $\mathcal{E}^{\text{external}}$. We get $H^1 \in \mathbb{R}^{C \times d}$ and $H^2 \in \mathbb{R}^{C \times D}$ through each GCN layer respectively. C and d denote the number of objects and the dimension of the representation, respectively. D denotes the depth of visual features produced by GCN.

To project $H^2 \in \mathbb{R}^{C\times D}$ and the visual features $X_{\text{visual}} \in \mathbb{R}^{D\times W\times H}$ to a common subspace, we adopt a feature mapping module. Specifically, we compressed the dimension of $X_{\text{visual}} \in \mathbb{R}^{D\times W\times H}$ to $\hat{X}_{\text{visual}} \in \mathbb{R}^{D\times (W\odot H)}$. Then, we transpose the dimension of $\hat{X}_{\text{visual}}$ to $\mathbb{R}^{(W\odot H)\times D}$. We further compressed channels of visual features with the same of the numbers of objects: $\hat{X}_{\text{visual}} \in \mathbb{R}^{(W\odot H)\times C}$ with two fully connection (FC) layers $F_1 \in \mathbb{R}^{D\times C}$. After that, we flatterned the multiple of H^2 and $\hat{X}_{\text{visual}}$ and got $\hat{K} \in \mathbb{R}^{\mathrm{D}\times\mathrm{W}\times\mathrm{H}}$. Furthermore, we concatenated X_{visual} and $\hat{K}$ and fed it to FC layer $F_2 \in \mathbb{R}^{2D\times D}$. In this way, we have converted the textual to the visual features.

However, despite the major challenge in trading off between annotation cost and noisy multi-modal pairs, common sense is not always recorded in text due to the reporting bias. Thus, we try to alleviate the bias with the intervention. With the same as the internal knowledge, we average mask of the i-th category features in the dimension D and got the deconfound external knowledge C_{external}. The overall logits are formulated by

$$C_{\text{external}} = \text{AVG}(F_2\{[\phi(F_1(\phi(X_{\text{visual}}))^T \odot H^2)]||X_{\text{visual}}\}^T) \quad (4)$$

Where $\phi(\cdot)$ denotes the dimension transpose function. $\odot, ||$ are matrix multiplication and matrix concatenation respectively.

3.3 Causal Intervention Module

Recalling the Fig.1, we get X's context (see in the red arrow) x and Y (see in the yellow arrow) after the image fed to the ViT encoder. The last layer of classification tasks is the Softmax layer: $P(y^c|x,c) = Softmax(f_y(x,c))$, where $f_y(\cdot)$ calculates the logits for N categories, and y denotes that $f(\cdot)$ is parameterized by Y's context y. The overall output of logits is defined as:

$$P(Y|do(X)) := \mathbb{E}_c[Softmax(f_y(x,c))] \quad (5)$$

We use the normalized weighted geometric mean (NWGM) [33] to move the outer expectation into the Softmax function as:

$$\mathbb{E}_c[Softmax(f(x,c))] \approx Softmax(\mathbb{E}_c[f_y(x,c)]) \quad (6)$$

For the classification task, we use the linear model $f_y(x,c) = \mathbf{W}_1 x + \mathbf{W}_2 \mathbb{E}[g_y(c)]$, where $\mathbf{W}_1, \mathbf{W}_2$ denote the fully connected layers, it is formulated by:

$$\mathbb{E}_c[f_y(x,c)] = \mathbf{W}_1 x + \mathbf{W}_2 \mathbb{E}_c[g_y(c)] \quad (7)$$

where $\mathbb{E}_c[g_y(c)]$ is calculated by the attention mechanism. Specifically, we are given $y \in Y$ and $c \in C$, the attention vector α is calculated by $softmax(q^T K/\sqrt{\sigma})$, then, we get $A = [\alpha; \ldots, \alpha]$ by the broadcasting operation. The most intuitive explanation is that we use attention mechanism to obtain the focus point between two objects, where $[;]$ denotes broacasting along the row. $q = W_3 y, K = W_4 C^T$. W_3, W_4 map each vector to the common subspace and σ is a constant scaling factor with the first dimension of W_3, W_4. Finally, $\mathbb{E}[g_y(c)] = \sum_c [A \odot C] P(c)$, where $\odot$ and $P(c)$ denote the element-wise product and prior statistic probability respectively. In summary, we obtain regions of interest similar to human visual system from a global perspective.

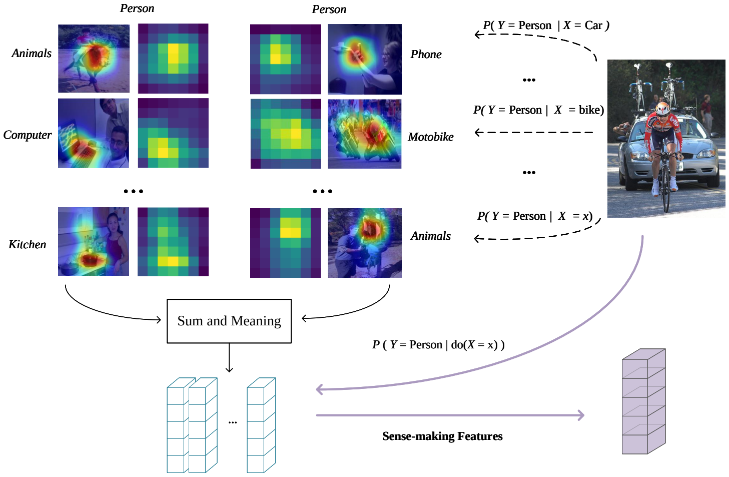

Fig. 3: “sense-making" processing.

3.4 Objective function

Given the two features, $x \in \mathbf{X}$ and $y \in \mathbf{Y}$ from ViT encoder, x and the context C are fed to the NWGM module, and the removed confounder features are obtained. Furthermore, we adapt the fully connected layer to learned the relationship p_i between y and $do(x)$. The loss of this processing is $\frac{1}{K}\sum_i \mathcal{L}_{cxt}(p_i, y_i^c)$, Finally an enhanced feature is feed to the decoder. Our training objective is formulated by:

$$\arg\min \mathcal{L} = \mathcal{L}_{seg}(p_i, \hat{y}_i^c) + \frac{1}{K}\sum_i \mathcal{L}_{cxt}(p_i, y_i^c) \tag{8}$$

where $\hat{y}_i^c$ denotes the ground-truth label provided by dataset, and y_i^c denotes the label on sub classification task. According to $P(Y|do(X))$, Y is one of the K context objects with the label y_i^c. $\mathcal{L}_{cxt}(p_i, y_i^c)$ is calculated by $-\log(p_i[y_i^c])$.

4 Experiments

We conduct experiments on semantic segmentation task semantic segmentation with Cityscapes and ADE20K. The details are as below.

4.1 Settings

Datasets: We use two commonly used datasets for semantic segmentation: Cityscapes [8], ADE20K [39]. The Cityscapes dataset contains 5000 images of driving scenes in urban environments (2975 for train, 500 for validation, 1525 for test). The resolution per image is 1024 × 2048 contains 19 categories of fine-grained annotations. ADE20K contains over 25K images (20k for training, 2k for validation, 3k for test). These images are densely annotated with an open dictionary label set.

Metric : We use mean intersection-over-union (mIoU) to calculate the ratio of the intersection and union of the two sets of true and predicted values. The classification task returns a true positive (TP), false positive (FP), true negative (TN) and false-negative (FN). It is formulated by

$$\begin{aligned}\mathrm{mIoU} &= \frac{\mathrm{TP}}{\mathrm{FP}+\mathrm{FN}+\mathrm{TP}} \\ &= \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}+\sum_{j=0}^{k} p_{ji}-p_{ii}}\end{aligned} \tag{9}$$

Where p_{ij} represents the total number whose true value is i but predicted to be j. p_{ii} denotes the number of TP, p_{ij} and p_{ji} denote FP and FN, respectively. $k+1$ is the count of classes (including the background class.)

Implementation details : We use the mmsegmentation toolbox to carry on experiments effectively. (i) the random resize with a ratio between 0.5 and 2, the random cropping $(768, 512)$ and 480 for Cityscapes, ADE20K, and the random horizontal flipping during training all the experiments. (ii) The total iteration is set to 80,000 and 160,000 for the experiments on Cityscapes and ADE20K, respectively, and both cases with batch size 16 and 8, respectively. (iii) We adopt a polynomial learning rate decay schedule and employ SGD as the optimizer. Momentum and weight decay are set to 0.9 and 0, respectively, for all the experiments on the two datasets. The initial learning rates are set to 0.001 on ADE20K and 0.01 on Cityscapes. (iv) To obtain the context C, we employed the pre-trained ViT model with the ground-truth labels as the input to extract the features for each object.

4.2 Comparision to state-of-the-art

We conducted comparative experiments on some representative models; the results are shown in Table 1. APNB [40], CCNET [13], SPNet [12] proposed to explore the way of enhancing the ability of spatial contextual representation. SETR [38], Segmenter [28] provides us with a new perspective, that is, using transformer-based encoder [18] to capture richer and more effective global context semantic information. Swin Transformer provides an effective patch embedding method based on shift windows to reduce network size. However, all of the above models do not provide a causal explanation. It is worth noting that we only plug the causal intervention module with the internal and external context in SETR and achieve a new SOTA in both Cityscapes and ADE20K.

Table 1: Comparison with the state-of-the-art methods.

Method	**Publication**	**Backbone**	**mIoU** (%)	
			Cityscapes	ADE20K
APNB	ICCV'19	ResNet101	81.30	45.24
CCNET	ICCV'19	ResNet101	81.40	45.76
SPNet	CVPR'20	ResNet101	82.00	45.60
EfficientFCN	ECCV'20	ResNet101	-	45.28
KRNet	ICASSP'21	ResNet101	82.20	45.65
SETR	CVPR'21	ViT-L	82.15	50.28
Swin Transformer	ICCV'21	Swin-L	-	53.50
Segformer	NIPS'21	Seg-L-Mask/16	82.20	51.80
Segmenter	ICCV'21	ViT-L	-	53.63
Ours	-	ViT-L	**83.21**	**54.48**

Qualitative Analysis Qualitative results are shown in Fig. 4. We use different coloured boxes to mark the differences between our model and the SETR. From the figure, we can observe that our proposed model has more accurate and more fine segmentation performance (marked with orange or yellow boxes). More objects are segmented: small, occluded, and indistinct segments. Global context information can effectively learn the co-occurrence relationship between different objects. However, some examples of objects being misclassified are circled with black boxes. It is effective for using causal intervention to remove the bias in contextual information.

4.3 Ablation Study

The contributions of different module: To fairly evaluate our proposed method, we carry out different settings and report results in table 2. Multi-scale test with random flipping (MS+Flip) is commonly used to improve semantic segmentation performance. We plugged the internal contextual information, removed confounder bias in SETR, and achieved 52.35% mIoU. It significantly increases mAP over the baseline by up to 2.07. By using MS+Filp, we achieved 53. 28% mIoU. Furthermore, we study the influence of introducing external contextual information with two settings: 1) The external knowledge without de-confounding (marked by **External**); 2) The external knowledge with de-confounding (marked by **External***). We first introduced the external knowledge without de-confounding, and got the margin improvement (0/13%mIoU).

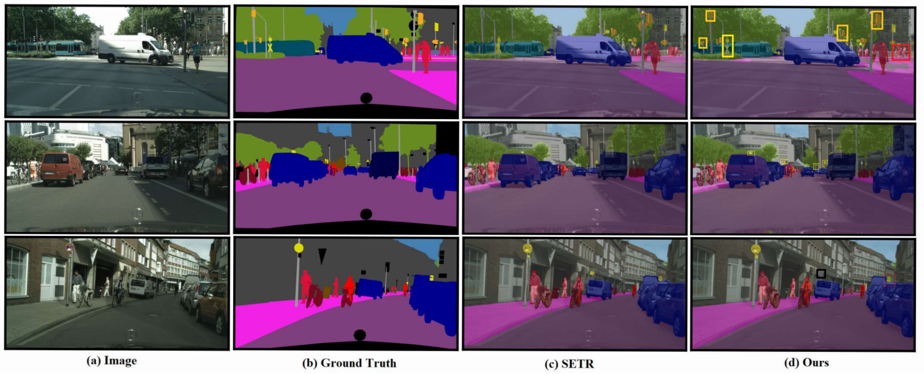
(a) Image (b) Ground Truth (c) SETR (d) Ours

Fig. 4: Qualitative results on Cityscapes dataset.

Secondly, with de-confounded external knowledge, we achieve an improvement of 1.05%mIoU.Plugging the de-confounded external knowledge without MS+Flip, the mIoU is up to 52. 70%. Finally, we adopt MS+Flip, internal knowledge, and de-confounded external knowledge. The overall performance is over SETR by up to 8%. In summary, it is effective to improve the performance of ViT by introducing internal knowledge and external knowledge.

Table 2: Ablation study on ADE20K datasets. **Internal**: intervention with internal knowledge. **External**: The external knowledge without de-confounding. **External***: The external knowledge with de-confounding

Method	**MS+Filp**	**Internal**	**External**	**External***	**mIoU** (%)
SETR	✓				50.28
Ours		✓			52.35
Ours	✓	✓			53.28
Ours			✓		50.41
Ours	✓		✓		50.94
Ours				✓	51.33
Ours	✓			✓	52.70
Ours		✓	✓		52.93
Ours	✓	✓	✓		53.40
Ours		✓		✓	53.93
Ours	✓	✓		✓	54.48

The influence of adapting different GCN layers: As mentioned in Section 3.2, we map the natural language co-occurrence probability in the Visual Gnome dataset to a common feature space using the word embedding method.Then, the GCN is used to extract the semantic information of the different objects from the common feature space. Therefore, it is necessary to discuss how many GCN layers are most beneficial for our method. As shown in Table.3, when the number of GCN layers increases, segmentation performance drops on both datasets. The optimal layer number of GCN is related to the sparsity degree of the adjacency matrix. When the sparsity degree of the graph is low, the over-smoothing phenomenon will soon occur. As a result, performance degradation occurs when more GCN layers are used.

Table 3: The influence of different depths of GCN in external knowledge mapping.

Layers	**Encoder**	**Cityscapes**	**ADE20K**	**mIoU (%)**
2 layers	ViT-L	✓		**81.58**
2 layers	ViT-L		✓	**52.70**
3 layers	ViT-L	✓		79.13
3 layers	ViT-L		✓	51.17
4 layers	ViT-L	✓		79.01
4 layers	ViT-L		✓	50.85
5 layers	ViT-L	✓		78.71
5 layers	ViT-L		✓	50.58

The influence of different word embedding methods: Similar to Section.4.3, we use different word embedding methods to integrate external knowledge better. Thus, we investigate four different word embedding methods, including Word2vec, GoogleNews, GloVe and the FastText word embedding. Fig.5 shows the results using different word embeddings on Cityscapes and ADE20K. From the figure, we can see that when using different word embeddings as graph's nodes, the segmentation mIoU will not be affected significantly. Furthermore, using GloVe could lead to better performance. The reason is that the word embeddings learned from large text corpus maintain some implicit knowledge.

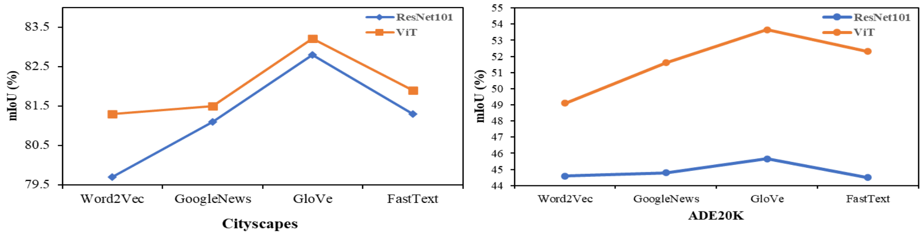

Fig. 5: Influence of word embedding methods

4.4 Plug and Play

We chose several representative ViT-based encoders and plugged our method into them to test its extensibility. For each method, we select the backbone encoder that performs best on the ADE20K benchmark. This also implies that they occupy the greatest number of parameters. For more details, we show in table 4. We plug our method in SETR (the baseline) and achieve 54.48 % mIoU. It is up to 4.20 % higher than SETR. Swin-Large [18], Seg-L-Mask/16 [28] is the advanced version of ViT. Swin-Large introduces the W-MSA operation, which reduces computation complexity. They use windows instead of patches, which makes the computational complexity of W-MSA linear with image size. Seg-L-Mask/16 adapts image patch join processing and class embedding for remodeling the global context. Furthermore, the Mask Transformer can perform direct segmentation rather than class embedding. By incorporating our method, the mIoU is increase to 0.68 % and 1.33 %, respectively. SegFormer contains a novel hierarchical Transformer encoder as well as a lightweight All-MLP decoder. It generates multi-scale features that do not require position coding, thereby avoiding position-coding interpolation, which leads to performance degradation when the test resolution differs from the training resolution. By plugging our method into Segformer, we achieveWe achieve 53.71 % mIoU (up to 1.46 % higher than SegFormer). It is worth noting that our method introduces extra parameters with calculating scale dot-product attention ($2 \times 512 \times 1024$), linear addition ($2 \times N \times 1024$) and feature embedding ($N \times 1024$). N denotes the count of categories. In short, we improved on several benchmarks by adjusting a few parameters. This demonstrates that our method works and that it can be applied to any other ViT-based segmentation encoder.

Table 4: Plug our method in different ViT-based Methods.

Method	**Backbone**	**Params** (M)	**Original mIoU** (%)	**mIoU** (%)
SETR	ViT-large	308	50.30	54.48
Swin-Transformer	Swin-large	234	53.53	54.21
SegFormer	MiT-B5	84	51.80	53.13
Segmenter	Seg-L-Mask/16	307	52.25	53.71

5 Conclusions

We provide a rethinking to semantic segmentation based on vision transformer model design and contribute a causal intervention attempt. Different from other tasks, we explained the model based on causal intervention. Using only feature concatenation, we improve on segmentation task, and then the model is closer to the human recognition system. Furthermore, causality can not only be explained by intervention but also many,counterfactual methods deserve further consideration. Therefore, we will further use causality to explore the next generation of artificial intelligence in the future.

References

1. Badde, S., Hong, F., Landy, M.S.: Causal inference and the evolution of opposite neurons. Proceedings of the National Academy of Sciences **118**(36) (2021)
2. Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., Goyal, A., Pal, C.: A meta-transfer objective for learning to disentangle causal mechanisms. arXiv preprint arXiv:1901.10912 (2019)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020)
4. Chalupka, K., Perona, P., Eberhardt, F.: Visual causal feature learning. arXiv preprint arXiv:1412.2309 (2014)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
6. Chen, S., Li, Z., Tang, Z.: Relation r-cnn: A graph based relation-aware network for object detection. IEEE Signal Processing Letters **27**, 1680–1684 (2020)
7. Chen, S., Li, Z., Yang, X.: Knowledge reasoning for semantic segmentation. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2340–2344 (2021)
8. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3213–3223 (2016)
9. Dasgupta, I., Wang, J., Chiappa, S., Mitrovic, J., Ortega, P., Raposo, D., Hughes, E., Battaglia, P., Botvinick, M., Kurth-Nelson, Z.: Causal reasoning from meta-reinforcement learning. arXiv preprint arXiv:1901.08162 (2019)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3146–3154 (2019)
12. Hou, Q., Zhang, L., Cheng, M.M., Feng, J.: Strip pooling: Rethinking spatial pooling for scene parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4003–4012 (2020)
13. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 603–612 (2019)
14. Kalainathan, D., Goudet, O., Guyon, I., Lopez-Paz, D., Sebag, M.: Sam: Structural agnostic model, causal discovery and penalized adversarial learning (2018)
15. Kocaoglu, M., Snyder, C., Dimakis, A.G., Vishwanath, S.: Causalgan: Learning causal implicit generative models with adversarial training. arXiv preprint arXiv:1709.02023 (2017)
16. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)

17. Li, Z., Sun, Y., Zhu, J., Tang, S., Zhang, C., Ma, H.: Improve relation extraction with dual attention-guided graph convolutional networks. Neural Computing and Applications **33**(6), 1773–1784 (2021)
18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
19. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
20. Lopez-Paz, D., Nishihara, R., Chintala, S., Scholkopf, B., Bottou, L.: Discovering causal signals in images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6979–6987 (2017)
21. Pearl, J.: Theoretical impediments to machine learning with seven sparks from the causal revolution. arXiv preprint arXiv:1801.04016 (2018)
22. Pearl, J., Glymour, M., Jewell, N.P.: Causal inference in statistics: A primer. John Wiley & Sons (2016)
23. Pearl, J., Mackenzie, D.: The book of why: the new science of cause and effect. Basic books (2018)
24. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. pp. 1532–1543 (2014)
25. Quan, Y., Li, Z., Chen, S., Zhang, C., Ma, H.: Joint deep separable convolution network and border regression reinforcement for object detection. Neural Computing and Applications **33**(9), 4299–4314 (2021)
26. Redondo-Cabrera, C., Baptista-Ríos, M., López-Sastre, R.J.: Learning to exploit the prior network knowledge for weakly supervised semantic segmentation. IEEE Transactions on Image Processing **28**(7), 3649–3661 (2019)
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**, 91–99 (2015)
28. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. arXiv preprint arXiv:2105.05633 (2021)
29. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021)
30. Wang, T., Huang, J., Zhang, H., Sun, Q.: Visual commonsense representation learning via causal inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (June 2020)
31. Wei, H., Li, Z., Huang, F., Zhang, C., Ma, H., Shi, Z.: Integrating scene semantic knowledge into image captioning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **17**(2), 1–22 (2021)
32. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. arXiv preprint arXiv:2103.15808 (2021)
33. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057 (2015)
34. Yang, X., Zhang, H., Qi, G., Cai, J.: Causal attention for vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9847–9857 (2021)

35. Zhang, D., Zhang, H., Tang, J., Hua, X.S., Sun, Q.: Causal intervention for weakly-supervised semantic segmentation. Advances in Neural Information Processing Systems **33** (2020)
36. Zhang, H., Zhang, H., Wang, C., Xie, J.: Co-occurrent features in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 548–557 (2019)
37. Zhang, J., Li, Z., Zhang, C., Ma, H.: Stable self-attention adversarial learning for semi-supervised semantic image segmentation. Journal of Visual Communication and Image Representation **78**, 103170 (2021)
38. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6881–6890 (2021)
39. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 633–641 (2017)
40. Zhu, Z., Xu, M., Bai, S., Huang, T., Bai, X.: Asymmetric non-local neural networks for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 593–602 (2019)