KE-GAN: Knowledge Embedded Generative Adversarial Networks for Semi-Supervised Scene Parsing

Mengshi Qi[1,2], Yunhong Wang*[1,2], Jie Qin[3], and Annan Li[2]

[1]State Key Laboratory of Virtual Reality Technology and Systems
School of Computer Science and Engineering, Beihang University, China
[2]Beijing Advanced Innovation Center for Big Data and Brain Computing
[3]Inception Institute of Artificial Intelligence, UAE

Abstract

In recent years, scene parsing has captured increasing attention in computer vision. Previous works have demonstrated promising performance in this task. However, they mainly utilize holistic features, whilst neglecting the rich semantic knowledge and inter-object relationships in the scene. In addition, these methods usually require a large number of pixel-level annotations, which is too expensive in practice. In this paper, we propose a novel Knowledge Embedded Generative Adversarial Networks, dubbed as KE-GAN, to tackle the challenging problem in a semi-supervised fashion. KE-GAN captures semantic consistencies of different categories by devising a Knowledge Graph from the large-scale text corpus. In addition to readily-available unlabeled data, we generate synthetic images to unveil rich structural information underlying the images. Moreover, a pyramid architecture is incorporated into the discriminator to acquire multi-scale contextual information for better parsing results. Extensive experimental results on four standard benchmarks demonstrate that KE-GAN is capable of improving semantic consistencies and learning better representations for scene parsing, resulting in the state-of-the-art performance.

1. Introduction

Scene parsing [22, 23, 42, 45, 47], *i.e.* assigning semantic class labels to pixels in a scene image, is one of the most crucial research topics in computer vision. It has a variety of applications, such as autonomous driving and robot navigation. As shown in Figure 1, every pixel annotated in the image can be classified into two categories, *i.e.* stuff (*e.g.* sky, field, sea, and beach) and object (*e.g.* tree, people, and dog). Because of the diversity of stuff/objects and complex relationships between them, scene parsing still remains a very challenging problem.

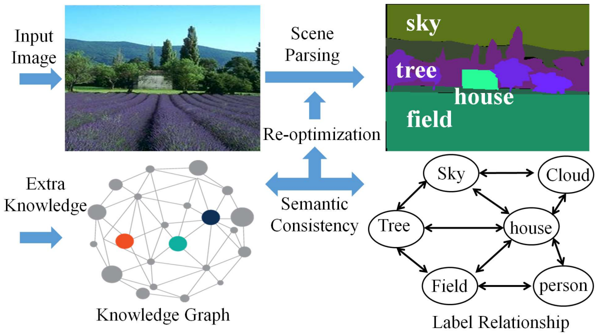


Figure 1. Scene parsing with knowledge embedding. We measure the semantic consistency between class labels to re-optimize the parsing result, by utilizing a large-scale knowledge graph to quantify the relationships between labels in the scene image.

A variety of significant efforts have been devoted to this issue in the past decades. The early attempts mainly include non-parametric methods [40] and parametric ones [19, 37]. However, most of them work in a similar manner as coarse-level image retrieval, making them arduous to measure accurate similarities between images and pixels. Meanwhile, these methods are lack of generality due to employing low-level hand-crafted features for classification. Recently, most of the state-of-the-art approaches [6, 7, 26, 44] adopt Convolutional Neural Network (CNN) based models for scene parsing and have achieved very promising results. However, the above CNN based approaches tend to predict all semantic labels independently from each other, overlooking the semantic relationship between each label and pixel. This directly leads to some failure cases including inaccurate confusion categories, missing inconspicuous classes,

*Corresponding author: yhwang@buaa.edu.cn and mismatching relationships between labels. Furthermore, these methods often utilize several time-consuming post-processing procedures to enforce the spatial contiguity in the output segmentation map, such as Conditional Random Field (CRF) [1, 48]. In addition, they require large-scale fully pixel-level annotated training data, which is too expensive to obtain.

In this paper, we attempt to tackle the above problems in a semi-supervised manner, by exploiting the underlying latent structures from unlabeled and synthesized examples. Furthermore, since pixel-level annotations and inter-label relationships are not invariably available, we attempt to guide the training process with some extra knowledge to capture semantic correlations between different classes. Specifically, apart from unlabeled real data, we aim to generate plausible scene images using the Generative Adversarial Networks (GAN) [15]. The generated images can complement the unlabeled real images to provide more meaningful structured features and benefit the subsequent pixel-level classification task. In terms of the extra knowledge, a typical way to provide the semantic domain knowledge is using *Knowledge Graph* [33], which consists of nodes and edges for representing concepts and inter-concept relationships, respectively.

Based on the above observations, we propose a novel GAN-based semi-supervised model for scene parsing, namely *Knowledge Embedded Generative Adversarial Networks (KE-GAN)*. As depicted in Figure 2, our KE-GAN adopts a deconvolution-based generator to create visual data with only noise as the auxiliary data, and employs a discriminator with the Fully Convolutional Networks [26] to parse scene images. Moreover, a pyramid architecture is incorporated to estimate multi-scale contextual information in a fine-grained way. The auxiliary semantic knowledge is derived from the *Knowledge Graph* and embedded into the KE-GAN, which enforces high-order semantic consistencies of labels in a scene image. Through the adversarial training, we optimize a joint objective function that consists of a conventional multi-class cross-entropy loss with an adversarial term, and a knowledge relation loss. Experimental results on four datasets demonstrate that our KE-GAN is beneficial for learning meaningful contextual features and complex relationships between different labels for accurate pixel-level classification.

Our main contributions are summarized as follows:

- We propose a novel Generative Adversarial Network based framework, *i.e*. KE-GAN, for scene parsing in a semi-supervised manner. KE-GAN generates extra scene images for training and captures multi-scale contextual features with a pyramid pooling module.
- We develop a knowledge graph guided optimization strategy during the adversarial training, which incorporates extra semantic knowledge into scene parsing and improves semantic consistencies between different labels.
- Extensive experiments on four public benchmarks comprehensively verify the superior performance of the proposed KE-GAN compared to the state-of-the-art methods.

2. Related Work

Semantic Segmentation: Traditional methods [1, 37, 40, 48] utilized graph structures to extract contextual information of images, *e.g*. Markov Random Field (MRF) or Conditional Random Field (CRF). Afterwards, deep learning based approaches [14, 21, 22, 23, 42, 43, 45, 47] have been studied in a plethora of works. A multi-scale convolutional network was introduced to label each pixel in [13]. Fully Convolutional Networks (FCN) [26] mapped the input RGB space to a semantic pixel space by an end-to-end training process with up-sampling filters. Dilated convolutions [26, 44] were adopted to systematically aggregate multi-scale contextual information without losing resolution. U-net [36] utilized skip connections for biomedical image segmentation. Semi-supervised semantic segmentation was tackled with deep neural networks in [18, 31]. However, they both focused on object segmentation and neglect the semantic knowledge and complex relationships in scene images, which is the main motivation of our KE-GAN.

Knowledge Representation: The knowledge in the real world can be represented as a graph structure where each node and edge represent one entity/concept and relationship respectively. RESCAL [30] was one of the earliest knowledge graph embedding models based on matrix factorization. TRANSE [3] introduced knowledge graph into translation embedding. Knowledge Vault [10] was proposed to extract web-scale probabilistic knowledge repositories from Web content. Fang et al. [12] introduced knowledge graph into object detection. Two closely related works to ours are [22, 45]. Liang *et al.* [22] adopted a semantic neuron graph to incorporate the semantic label information. Zhao *et al.* [45] employed word embedding for scene parsing. However, the above two works leveraged the annotated labels as the extra knowledge, lacking practicality for general applications. In contrast, our proposed KE-GAN employs a widely-adopted large-scale *knowledge graph* into scene parsing.

Adversarial Learning: Recently, Generative Adversarial Networks (GANs) have exhibited remarkable capabilities in image generation [9, 15, 32] and representation learning [11, 15, 20, 39]. Briefly, GANs are composed of a generator for synthesizing images that are difficult to distinguish from real images, and a discrimina-

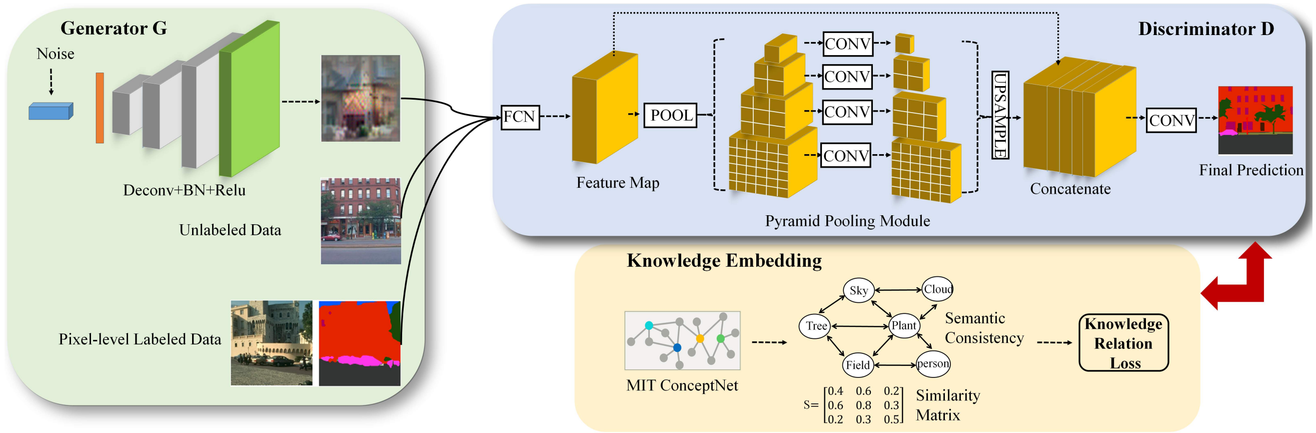


Figure 2. Overview of the proposed KE-GAN. A deconvolution based generator G constructs the extra scene images only with the noise input. The discriminator accepts the generated data, unlabeled data and pixel-level annotated data as the input to learn class-level confidence maps for each semantic label. Moreover, we employ Fully Convolutional Networks (FCN) to obtain the feature map, and apply a pyramid pooling module to extract multi-scale rich contextual representations. In addition, knowledge embedding is performed with a knowledge relation loss to re-optimize the generated scene parsing result, by utilizing a similarity matrix learned from the knowledge graph *MIT ConceptNet*.

tor that is optimized to discriminate real images from generated ones. Deep Convolutional Generative Adversarial Networks (DCGANs) [34] were introduced in an unsupervised way to construct images by up-sampling. Conditional GANs (cGANs) were firstly introduced for image tag prediction [29], and then applied in video prediction [28], text to image synthesis [35], and image-to-image translation [20, 50]. However, limited works have exploited GANs for semi-supervised scene parsing. Luc et al. [27] presented a generator to segment a given image into probability maps through FCN, and the discriminator was used to distinguish generated maps from ground truth. Souly et al. [38] proposed semi- and weakly-supervised semantic segmentation using GANs. However, our proposed KE-GAN incorporates extra domain knowledge to capture semantic correlations between different classes for better semi-supervised scene parsing performance.

3. Proposed Approach

3.1. Overview

As illustrated in Figure 2, our KE-GAN consists of a generator and a discriminator. The generator G is a deconvolution network, which is frequently used for image generation. The discriminator D serves as a conventional scene parser, which adopts Fully Convolutional Networks with a pyramid pooling module. The input of the discriminator includes the generated data, unlabeled data, and pixel-level annotated data. The output of the discriminator is the class-level confidence maps for each semantic label of the pixel. The discriminator attempts to suppress the probability of real classes for the pixel of the generated image and encourage high confidence of semantic labels for the pixel of the unlabeled image through adversarial training. By virtue of this framework, we can embed extra semantic knowledge into the adversarial learning process. As a result, a better knowledge-aware scene parser can be created based on the framework. Particularly, the knowledge embedded in the KE-GAN is derived from the MIT ConceptNet dataset [25]. A novel knowledge relation loss is utilized to re-optimize the parsing results between neighbor pixels, by taking the spatial positions and semantic relationships between pixels into consideration. It is noteworthy that during the test we only use the discriminator network as the scene parser. We adopt the softmax layer of the discriminator to classify each pixel into semantic classes by outputting a set of class-specific probability maps. Finally, we assign the label that has the highest possibility to the corresponding pixel.

3.2. A Brief Review of GANs

Generally, GANs [15] contain a generator G and a discriminator D, which are iteratively optimized in an adversarial way. We follow the adversarial training process, *i.e.* the generator G tries to model the underlying data distribution and confuse the discriminator D, and D aims at distinguishing the fake samples generated by G from the ground truth (*i.e.* true distribution $p_{data}(x)$). In other words, G and D are competitors in a min-max game. The loss function of optimizing GAN is formulated as follows:

$$\begin{aligned}\min_G \max_D \mathcal{V}(D,G) =& \mathbb{E}_{x\sim p_{data}(x)}[\log(D(x))] + \\ & \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))],\end{aligned} \tag{1}$$

where $\mathbb{E}$ is the empirical estimate of the probability. A noise variable z from distribution p_z is fed into G, and the output is denoted as $G(z)$. Distribution p_z is expected to converge to distribution p_{data}. Minimizing $\log(1 - D(G(z)))$ is equivalent to maximizing $\log(D(G(z)))$.

3.3. Generator

Synthesizing scene images only needs a forward pass through the generator G. The generator network of KE-GAN contains four deconvolution layers that transform noise into images, as manifested in Figure 2. In particular, all the deconvolution layers are followed by batch normalization (BN) and rectified linear units (ReLU) based non-linear activation. The noise term z is a 100-dimensional vector sampled from a uniform distribution $p_z(z)$. Then, we expect the synthetic image to resemble samples from the real data distribution, and the generator loss is formulated as the following:

$$\mathcal{L}_{GAN_G} = \mathbb{E}_{z \sim p_z(z)}[log(1 - D(G(z)))]. \tag{2}$$

3.4. Discriminator

The discriminative network D is proposed as a fully convolutional network [26, 38] with a pyramid pooling module to classify each pixel. Because of limited training data with pixel-level labels, we aim to leverage unlabeled data for improving the performance of the pixel classifier. To this end, the input of D includes three types of data: unlabeled scene images, images generated by G and scene images with pixel-level annotations. The discriminator is utilized to predict whether a pixel sample x fits the true distribution of data or not, and assign a label y to each pixel, where $y \in [1, ..., K]$, and K is the number of semantic classes. The x refers to the pixel in a scene image rather than an image. The discriminator is utilized to predict whether a pixel sample x fits the true distribution of data or not, and assigns a semantic label y to each pixel x or not. By adding fake samples as another category, the discriminator D will output $K + 1$ confidence maps for each pixel. D attempts to suppress the probability of real classes for generated data and encourage high confidence of semantic labels for unlabeled data. The discriminator loss is:

$$\begin{aligned}\mathcal{L}_{GAN_D} = &- \mathbb{E}_{x \sim p_{data}(x)}[\log(D(x))] \\ &- \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \\ &+ \gamma \mathbb{E}_{x,y \sim p(y,x)}[\mathcal{C}(y, P(y|x, D))],\end{aligned} \tag{3}$$

where

$$D(x) = [1 - P(y = fake|x)], \tag{4}$$

$\mathcal{C}$ denotes the cross-entropy loss between labels and predicted probabilities by $D(x)$, and $P(y|x)$ denotes the probability of assigning label y to pixel x. The first term in $\mathcal{L}_{GAN_D}$ is to reduce the possibility of assigning pixels to the fake class for unlabeled data, the second term is devised to distinguish fake samples generated by G from real data for D, and the third term forces all pixel to be classified accurately as one of the K categories for labeled data.

3.5. Multi-Level Pyramid Pooling

Inspired by multi-scale representation in deep learning [16, 46], it is rewarding to incorporate the information from different sub-regions to construct the global context. Thus, we introduce a pyramid pooling module as part of the discriminator, which is capable of extracting global contextual information effectively. As illustrated in Figure 2, we employ four pyramid scales in our model in a coarse-to-fine way, of which bin sizes are 1×1, 2×2, 3×3, and 6×6, respectively. The coarser level indicates the output of the global pooling of a single bin. In the next level, the whole feature map is divided into multiple sub-regions, and the pooled features for different locations are constructed. We adopt the 1×1 convolution layer to reduce the dimension of our contextual representation. As such, the pyramid size in the N-th level is reduced to $1/N$. Then, we upsample the low-dimensional feature maps to maintain the same size via bilinear interpolation. Finally, we concatenate representations from different levels into the final global representation.

In practice, the discriminator of KE-GAN extracts feature maps via utilizing a pre-trained ResNet model [17]. Subsequently, we adopt the Fully Convolutional Networks with the 4-level pyramid pooling module to fuse multi-level representations as the global prior information. Finally, we concatenate the prior with the original feature map, which is followed by a convolution layer to generate the final prediction map.

3.6. Knowledge Graph Embedding

As we mentioned previously, one core issue of scene parsing is how to preserve the semantic consistency. However, the complex relationships between different semantic labels are arduous to capture from limited training data. The *Knowledge Graph* [33] can capture millions of concepts or entities and their complex relationships. Thus, it is beneficial to adopting the extra knowledge from a large-scale knowledge graph to measure the similarity and extract the relationship between a pair of semantic labels. In the knowledge graph, each label can be denoted as a node and the relationship between different labels can be an edge. Hence, we can create a chain of relationships between labels. As an example, "computer" and "chair" are not directly connected, but we can find a chain of edges "computer above desk" and "desk beside chair" to indicate that they still have some semantic consistency.

To preserve the semantic consistency in a scene, we employ a crowd-sourced knowledge graph, *i.e. MIT ConceptNet* [25], which includes more than 4 million concepts and 9 million relationships. We also incorporate a constraint for measuring the similarity between labels, as depicted in Figure 2. Given a set of pre-defined labels (or concepts) $L = l_1, l_2, ..., l_n$, we formulate S as an $L \times L$ similarity symmetric matrix to record the degree of semantic consistency between labels l and l', $\forall(l, l') \in L$. Thus, the matrix S can model the prior knowledge and relationships between different labels. To construct S, we follow [12] and employ random walk with restart [41] to quantify the semantic consistency on a knowledge graph. The random walk is a sequence of nodes $(v_0, v_1, ..., v_t)$, and $p(v_t = l'|v_0 = l; \alpha)$ means that the probability of reaching concept l' in t steps if it starts from l, where α means the random probability of restart random walk. Through computing the probability, we formulate $R_{i,j}$ as the final higher probability to imply that we can find shorter paths from node i to j, also suggesting that the semantic consistency $S_{i,j}$ is higher. More details can be found in [41]. Consequently, the matrix S can be computed as follows:

$$\begin{aligned} R_{l,l'} &= \lim_{t\to\infty} p(r_t = l'|r_0 = l; \alpha), \\ S_{l,l'} &= S_{l',l} = \sqrt{R_{l,l'}R_{l',l}}. \end{aligned} \tag{5}$$

We define $\hat{y}_i$ as the semantic label predicted by the discriminator network and y_i as the ground truth at pixel i, respectively. Then, we formulate the knowledge relation loss between two pixels based on the similarity with Kullback-Leiber divergence:

$$\mathcal{L}_r^{i,j} = \begin{cases} D_{KL}(\hat{y}_j||\hat{y}_i) & \text{if } y_i = y_j, \\ \max\{0, M - D_{KL}(\hat{y}_j||\hat{y}_i) \cdot S_{\hat{y}_j,\hat{y}_i}\} & \text{otherwise,} \end{cases} \tag{6}$$

where the Kullback-Leiber divergence D_{KL} is formulated for two Bernoulli distribution P and Q with parameters p and q respectively: $D_{KL}(P||Q) = p \log p/q + \overline{p} \log \overline{p}/\overline{q}$ for $p, q \in [0, 1]$. Specifically, the KL divergence will be minimized when the ground truth label of the pixel i equals to the one of its neighbor j; otherwise, the KL divergence will be maximized until margin M. The overall knowledge relation loss is the average one over all pixels, as the following:

$$\mathcal{L}_r = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathcal{L}_r^{i,j}, \tag{7}$$

where n and m are the number of all the pixels and their neighbors, respectively. Hence, the knowledge relation loss is employed to make the labels assigned to the neighbor pixels which are more similar based on the knowledge graph and generate more reasonable parsing results.

3.7. Joint Objective

Finally, the joint loss of our KE-GAN can be formulated as:

$$\begin{cases} \mathcal{L}_G = \mathcal{L}_{GAN_G}, \\ \mathcal{L}_D = \lambda_g \mathcal{L}_{GAN_D} + \lambda_r \mathcal{L}_r, \end{cases} \tag{8}$$

where λ_g and λ_r are the hyper-parameters used for training. We train the generator G and the discriminator D alternately until optimality. Finally, G generates scene images as extra training examples, and D becomes a reliable estimator and outputs the desired scene parsing results through adversarial training.

4. Experimental Results

In this section, we evaluate our framework in terms of scene parsing on four public datasets, *i.e.* **ADE20K**, **Cityscapes**, **SiftFlow** and **CamVid**. We mainly conduct two tasks in our experiments, *i.e. fully-supervised scene parsing* by adopting all of the labeled data for training, and *semi-supervised scene parsing* by utilizing a part of labeled data for training.

4.1. Experimental Settings

ADE20K [49] is the most challenging scene parsing dataset, which was also used in ImageNet scene parsing challenge 2016. There are up to 150 semantic classes and totally 1,038 image-level labels on this dataset. Following [49], we divide the dataset into 20K, 2K, and 3K images for training, validation, and test for fully supervised learning. As for semi-supervised learning, we employ 30% of the pixel-level annotated data and regard the rest as the unlabeled data.

Cityscapes [8] consists of 19 semantic categories. There are 50 videos with a high resolution of 2048×1024 in driving scenes collected from different European cities. Each annotated frame utilized in the training process is the 20th frame of a 30-frame snippet. Following [8], we split the whole dataset into three parts: *i.e.* 2,975 training samples, 500 validation samples, and 1,525 test samples w.r.t. fully supervised learning. In contrast, for semi-supervised training, we only employ 30% of the pixel-level annotated data, and the rest are considered as the unannotated images.

SiftFlow [24] contains 2,688 images in eight outdoor scenes with a resolution of 256×256, belonging to 33 semantic classes. We follow [24] and adopt 2,488 images for training and 200 images for test with respect to fully supervised training. While for semi-supervised training, we only adopt 50% of the labeled data and treat the rest as unlabeled data.

CamVid [4] is composed of over ten minutes' videos, including about 11K frames, of which 701 images with a resolution of 960×720 are pixel-level annotated. There are 32 semantic labels. Following [2], we choose 70% and 30% of the labeled images as the training set and the test set, respectively. In our experiments, we adopt the training set for fully supervised learning and all the unlabeled frames as the unlabeled data for semi-supervised learning.

Evaluation Metrics: Similar to most previous works, we adopt three standard evaluation metrics: *i.e.* per-class accuracy (CA), per-pixel accuracy (PA), and mean Intersection-over-Union (mIoU). Specifically, PA is defined

Table 1. Comparison results of our KE-GAN and several state-of-the-art scene parsing methods on two large-scale datasets. We use fully labeled training data for fully-supervised training and 30% labeled data for semi-supervised training. The best results are in bold.

Methods	ADE20K			Cityscapes		
	CA	PA	mIoU	CA	PA	mIoU
FCN [26]	40.3	71.3	29.4	34.4	85.7	66.0
SegNet [2]	31.1	71.0	21.6	41.4	87.2	57.0
DilatedNet [44]	44.6	73.5	32.3	42.0	86.5	67.1
DeepLab v2 [5]	46.6	75.8	33.9	42.6	86.4	70.4
CascadeNet [49]	48.3	74.5	34.9	N/A	N/A	N/A
PSPNet(Res-101) [46]	N/A	80.6	41.9	N/A	N/A	N/A
WiderNet [42]	N/A	81.1	**43.7**	N/A	N/A	**80.1**
RefineNet(Res-101) [23]	N/A	N/A	40.7	N/A	N/A	73.6
PSANet(Res-101) [47]	N/A	**81.5**	43.7	N/A	N/A	78.4
DSSPNNet(Res-101)Universal [22]	N/A	81.3	43.6	N/A	N/A	75.5
KE-GAN Fully-Supervised	**50.2**	80.5	37.1	**43.7**	**89.3**	75.3
KE-GAN Semi-Supervised	46.2	78.9	35.2	41.5	87.2	71.6

as the percentage of all correctly classified pixels, while CA is the average of all category-wise accuracies.

Compared Methods: We compare our method with several state-of-the-art approaches, including FCN [26], SegNet [2], DilatedNet [44], DeepLab v2 [5], the cascade model [49], PSPNet [46], WiderNet [42], RefineNet [23], PSANet [47], DSSPNNet [22] and Souly fully-supervised and semi-supervised methods [38]. In all the experiments, the parameter settings of the above-mentioned methods are adopted from the corresponding papers.

4.2. Implementation Details

All the implementations are based on the open source PyTorch[1] framework. All the framework is trained on a single NVIDIA 1080 Ti GPU. In the experiments, the mini-batch size is set to eight and the Adam optimizer is used for training the generator. The learning rate is set to 1×10^{-4}, and the weight decay factor is 0.5 for every 2,000 iterations. The discriminator is trained with the standard Adam optimizer with the learning rate of 1×10^{-6}, and the momentum and weight decay are set to 0.9 and 0.0001, respectively. Based on the cross-validation on the training data, the hyper-parameters in the loss function, *i.e.* λ_g and λ_r, are set to 0.1 and 0.001, respectively, and γ is set to 2 empirically. For the stability of knowledge graph embedding, we set the random walk restarting probability $\alpha = 0.15$, and the margin M in the knowledge relation loss is set to 3.0 for all the experiments. In practice, we randomly shuffle or interleave the labeled data, unlabeled data and generated data during the training process for semi-supervised learning, and continue the learning process after 3,000 iterations with only labeled data due to the stability of the model. Meanwhile, we average the experimental results with different seeds to ensure robust evaluations. We update the generator network and the discriminator network iteratively, and only employ the discriminator for outputting the scene parsing results in the test phase.

4.3. Results and Analysis

Results on ADE20K: Quantitative comparison results are shown in Table 1. In the fully supervised setting, our framework outperforms other state-of-the-art methods in terms of CA. KE-GAN with the semi-supervised setting achieves competitive performance with other fully supervised approaches, which implies that the proposed adversarial training scheme can compensate for the lack of annotated data and capture the global contextual information for better performance. By using knowledge graph embedding, our method using fully-supervised training significantly surpasses the strong baseline models (FCN, SegNet, and Deeplab v2) by 3% to 10% in terms of CA and mIoU, once again demonstrating the capability of our model in improving the classification of semantic labels. Note that the PA of our semi-supervised KE-GAN is less than 1.5% compared to the fully-supervised setting, which indicates that KE-GAN under semi-supervised training can learn distinctive representations of each semantic category for pixel-level classification given limited annotated data. Some qualitative results of our method on the ADE20K dataset are illustrated in Figure 3. As can be seen, the parsing results with adversarial training are much smoother, and the class probabilities regarding large areas are enhanced using our model.

Results on Cityscapes: The detailed results are reported in Table 1 and Figure 4. As we can see from the table, our method with fully-supervised and semi-supervised training achieve the best and comparable results to the state-of-the-art methods, respectively w.r.t CA and PA. It is demonstrated that the pyramid-structure GANs and knowledge graph embedding increase per-class and per-category accuracy. Meanwhile, our model with adversarial training and

[1] https://pytorch.org/.

Table 2. Performance comparisons of our method and the state-of-the-art approaches on the Cityscapes dataset w.r.t. mIoU with different amount of data. The best results are in bold.

Methods	1/8	1/4	1/2	Full
FCN [26]	N/A	N/A	N/A	66.0
Dilation [44]	N/A	N/A	N/A	67.1
Deeplab v2 [5]	N/A	N/A	N/A	70.4
Baseline+pyramid	62.6	65.5	69.9	72.6
Baseline+pyramid+$\mathcal{L}_{GAN_G/GAN_D}$	66.1	69.2	71.5	73.6
Baseline+pyramid+$\mathcal{L}_r$	63.7	66.3	70.3	72.9
Baseline+pyramid+$\mathcal{L}_{GAN_G/GAN_D}$**+**$\mathcal{L}_r$	**66.9**	**70.6**	**72.2**	**75.3**

knowledge embedding is especially beneficial for capturing the interactions among labels by generating realistic images shown in Figure 4, in addition to leveraging the contextual information.

Components Analysis on Cityscapes: As can be seen from Table 2, we conduct further experiments on Cityscapes to analyze and identify the effect of each component of our KE-GAN. We randomly sample $1/8$, $1/4$, and $1/2$ of the training data as the labeled data, and use the rest as the unlabeled data. Firstly, we adopt the FCN net as our baseline model, and we analyze how much the pyramid module contributes to scene parsing. We can see that the pyramid module boosts about 7% w.r.t. mIoU compared to the baseline, which demonstrates the effectiveness of multi-level representation extracted by pyramid pooling. Secondly, we introduce the adversarial loss $\mathcal{L}_{GAN_G/GAN_D}$ into the model, and it leads to consistent performance improvement on different amount of training data, *e.g*. from 2% to 4%, suggesting that our adversarial training scheme can encourage the parsing model to learn the structural information and distinguish representation from the distribution of ground truth data. Finally, we examine the effect of integrating the knowledge relation loss $\mathcal{L}_r$ into our model. Obviously, the whole KE-GAN framework achieves the best performance. Therefore, if the extra semantic knowledge from *Knowledge Graph* is not embedded into our KE-GAN, the generated confidence map and the pair-wise label relationship captured by the discriminator would be meaningless and inconsistent, deteriorating the final parsing performance. Moreover, we can observe that our *Baseline+pyramid* only with the adversarial loss achieves better performance than that only with knowledge embedding, demonstrating that $\mathcal{L}_{GAN_G/GAN_D}$ is more important in our KE-GAN. In addition, by using different numbers of labeled data for semi-supervised learning, our whole model still achieves 5% to 10% performance gain over fully-supervised FCN/Dilation/DeepLab v2. All the above observations indicate that semi-supervised learning is crucial for our model and each component of KE-GAN works effectively and complementarily.

Results on SiftFlow: The results on SiftFlow are depicted in Table 3. Generally, KE-GAN outperforms the state-of-the-art. In particular, by combining the adversarial training and the knowledge relation constraint, our model with semi-supervised learning obtains the best performance, improving 4% and 6% across all metrics compared to Souly Semi-Supervised and Fully-Supervised methods, respectively. Figure 5 exhibits a few qualitative results obtained using our approach. These results suggest that even the small objects (*e.g*. person, trail and different building) can be labeled correctly.

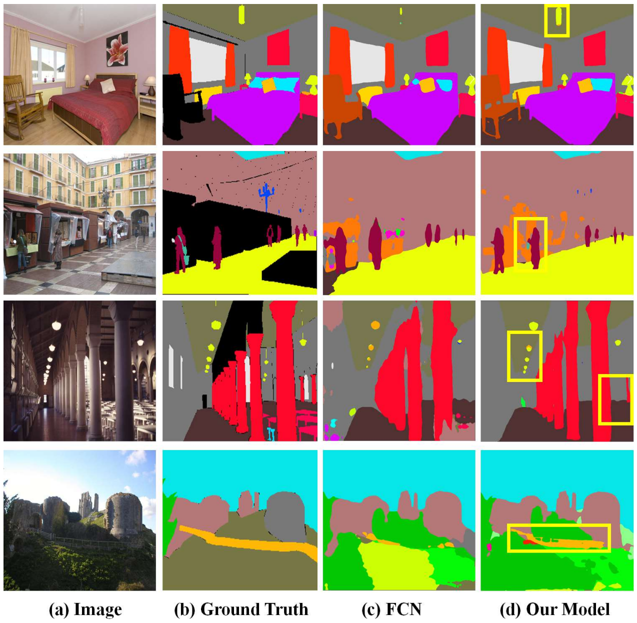

(a) Image (b) Ground Truth (c) FCN (d) Our Model

Figure 3. Qualitative parsing results on the ADE20K dataset. The improved labeled results by our KE-GAN are denoted in yellow box.

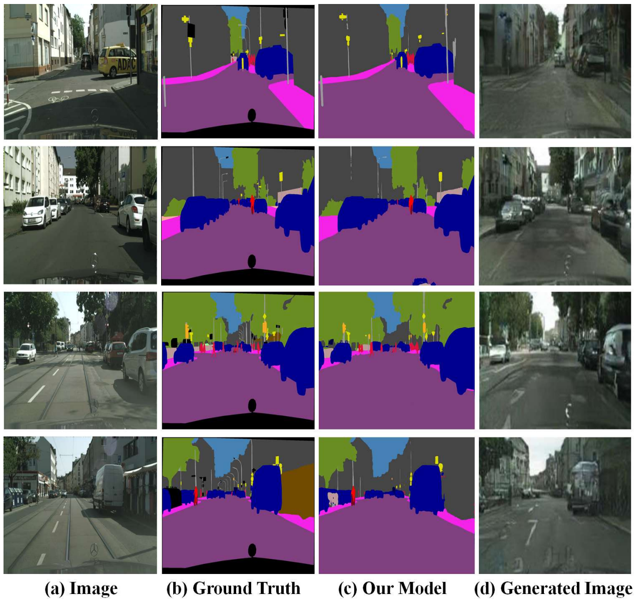
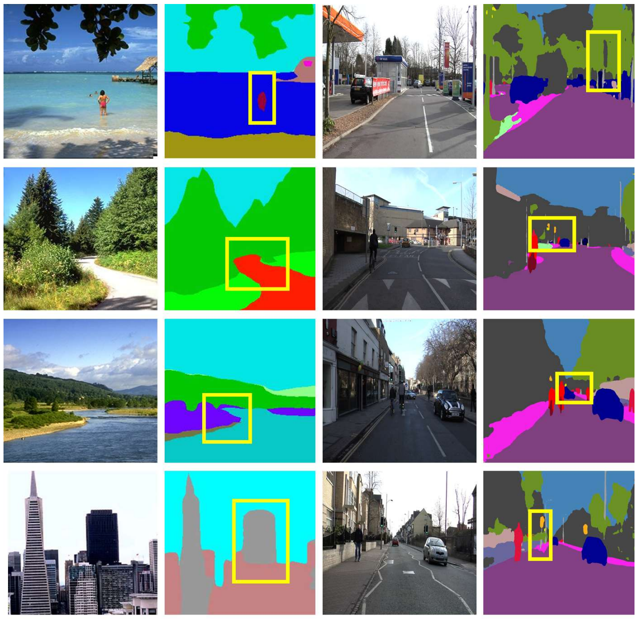

(a) Image (b) Ground Truth (c) Our Model (d) Generated Image

Figure 4. Qualitative parsing results and generated examples on the Cityscapes dataset.



(a) SiftFlow (b) CamVid

Figure 5. Qualitative parsing results on the SiftFlow and CamVid datasets. The improved labeled results by our KE-GAN are denoted in yellow box.

Results on CamVid: Finally, we evaluate our KE-GAN on the CamVid dataset. Table 4 denotes the quantitative results, where our KE-GAN with fully supervised learning obtains the best performance in terms of PA, improving the per-class accuracy from 2% to 7%. This indicates that our model is capable of recognizing more per pixel accurately. Moreover, KE-GAN with semi-supervised training achieves the best results w.r.t. CA and mIoU. The reason why our KE-GAN with semi-supervised learning can outperform that with fully-supervised learning is that our KE-GAN can generate more images as additional training examples for adversarial learning to recognize semantic classes accurately on such a small-scale dataset. From the qualitative results in Figure 5, we can see that our KE-GAN learns better hidden structure and knowledge relation, which contribute to the enhanced parsing results. Through knowledge embedding, our KE-GAN can also learn reasonable logistic relationships between objects, such as sky and mountains at the top, and buildings at the bottom. Moreover, a few small objects (*e.g*. pole, pedestrian or bicyclist) can be labeled correctly by employing extra data, which shows that introducing extra data generated by our KE-GAN is beneficial to refining the segmentation performance.

In summary, it is clear that the auxiliary data generated by adversarial training boosts the performance of scene parsing, where there is only limited labeled data. This is also the reason why semi-supervised KE-GAN can achieve better performance than that with full-supervised learning on small-scale datasets, *e.g*. CamVid and SiftFlow. Furthermore, the adversarial loss we proposed is able to learn more meaningful features for pixel-level classification. In addition, introducing the semantic consistency with our knowledge relation loss derived from the knowledge graph helps the discriminator to discover the relationships among labels, and improves category-level classification. Note that our model focuses on the task of semi-supervised scene parsing, and thus adopts a basic structure similar to FCN for segmentation. Our KE-GAN can be combined with the state-of-the-art segmentation models for further improved performance.

Table 3. Performance comparisons on the SiftFlow dataset with all labeled data for fully supervised learning and 50% of annotated data for semi-supervised learning. The best results are in bold.

Methods	PA	CA	mIoU
Souly Fully-Supervised [38]	79.0	28.3	21.0
Souly Semi-Supervised [38]	81.0	33.0	23.2
KE-GAN Fully-Supervised	83.2	36.1	25.9
KE-GAN Semi-Supervised	**85.3**	**37.6**	**27.2**

Table 4. Performance comparisons on the CamVid dataset using fully labeled training data and 1K unlabeled frames from the corresponding videos. The best results are in bold.

Methods	PA	CA	mIoU
SegNet-Basic [2]	82.2	62.3	46.3
SegNet-Pretrained [2]	88.6	65.9	50.2
DeepLab v2 [5]	84.6	62.6	61.6
Souly Fully-Supervised [38]	88.4	66.7	57.0
Souly Semi-Supervised [38]	87.0	72.4	58.2
KE-GAN Fully-Supervised	**89.2**	75.3	60.2
KE-GAN Semi-Supervised	87.9	**76.5**	**61.9**

5. Conclusion

In this paper, we propose a novel GANs based framework for semi-supervised scene parsing with knowledge embedding. The proposed KE-GAN generates scene images for data augmentation, deriving and quantifying semantic consistency with the help of a large-scale knowledge graph. For extracting rich contextual information from scene images, a pyramid pooling module is designed and integrated into the discriminator to segment scene images in pixel levels. Extensive experiments conducted on four widely-adopted datasets have indicated that our approach outperforms the state-of-the-art semi-supervised approaches and achieves competitive performance with fully-supervised methods.

6. Acknowledgment

This work was partly supported by the National Key Research and Development Plan of China under Grant 2016YFB1001002 and the Foundation for Innovative Research Groups through the National Natural Science Foundation of China under Grant 61421003.

References

[1] Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip HS Torr. Higher order conditional random fields in deep neural networks. In *ECCV*. Springer, 2016.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI*, 39(12):2481–2495, 2017.

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NeurIPS*, 2013.

[4] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*. Springer, 2008.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2018.

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*. Springer, 2018.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*. IEEE, 2016.

[9] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *NeurIPS*, 2015.

[10] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *SIGKDD*. ACM, 2014.

[11] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *NeurIPS*, 2016.

[12] Yuan Fang, Kingsley Kuan, Jie Lin, Cheston Tan, and Vijay Chandrasekhar. Object detection meets knowledge graphs. In *AAAI*, 2017.

[13] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *TPAMI*, 35(8):1915–1929, 2013.

[14] Weifeng Ge, Sibei Yang, and Yizhou Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *CVPR*. IEEE, 2018.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*. Springer, 2014.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*. IEEE, 2016.

[18] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *NeurIPS*, 2015.

[19] Wei-Chih Hung, Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Scene parsing with global context embedding. In *ICCV*. IEEE, 2017.

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*. IEEE, 2017.

[21] Shu Kong and Charless C Fowlkes. Recurrent scene parsing with perspective understanding in the loop. In *CVPR*. IEEE, 2018.

[22] Xiaodan Liang, Hongfei Zhou, and Eric Xing. Dynamic-structured semantic propagation network. In *CVPR*. IEEE, 2018.

[23] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*. IEEE, 2017.

[24] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *TPAMI*, 33(5):978–994, 2011.

[25] Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 2004.

[26] Jonathan Long, Evan Shelhamer, Trevor Darrell, et al. Fully convolutional networks for semantic segmentation. In *CVPR*. IEEE, 2015.

[27] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. *arXiv*, 2016.

[28] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv*, 2015.

[29] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv*, 2014.

[30] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *ICML*, 2011.

[31] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*. IEEE, 2015.

[32] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*. IEEE, 2016.

[33] Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 2017.

[34] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv*, 2015.

[35] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. 2016.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 2015.

[37] Nasim Souly and Mubarak Shah. Scene labeling using sparse precision matrix. In *CVPR*. IEEE, 2016.

[38] Nasim Souly, Concetto Spampinato, Mubarak Shah, et al. Semi and weakly supervised semantic segmentation using generative adversarial network. In *ICCV*. IEEE, 2017.

[39] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv*, 2016.

[40] Joseph Tighe, Marc Niethammer, and Svetlana Lazebnik. Scene parsing with object instances and occlusion ordering. In *CVPR*. IEEE, 2014.

[41] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Fast random walk with restart and its applications. In *ICDM*. IEEE, 2006.

[42] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *arXiv preprint arXiv:1611.10080*, 2016.

[43] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*. IEEE, 2018.

[44] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv*, 2015.

[45] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open vocabulary scene parsing. In *CVPR*. IEEE, 2017.

[46] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*. IEEE, 2017.

[47] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*. Springer, 2018.

[48] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *ICCV*. IEEE, 2015.

[49] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, pages 1–20, 2016.

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *ICCV*, 2017.