Recent trends in Diffusion models

2024년도 동계 세미나

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
Hanni Oh

Outline

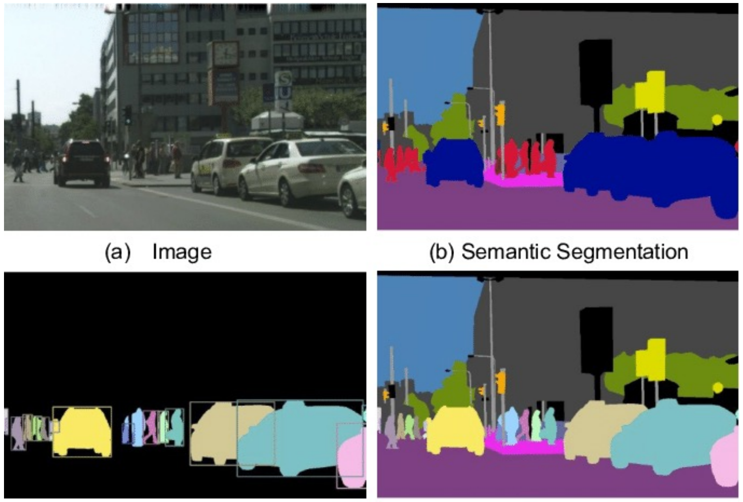
- Background
  - Open-vocabulary panoptic segmentation
  - Stable Diffusion

- Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models
  - CLIP Embedding
  - CVPR 2023 Highlight

- InstructPix2Pix: Learning To Follow Image Editing Instructions
  - CVPR 2023 Highlight

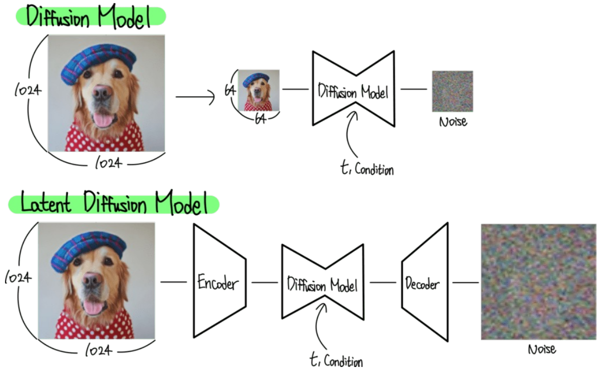
Background

- **Open-vocabulary panoptic segmentation**
  - Open-vocabulary
    - 일반적인 closed-vocabulary에서는 training dataset에 정의된 class만 감지
    - 그러나 open-vocabulary는 어떠한 사전 지식 없이 미리 정의되지 않은 class도 감지
    - 최종적으로는 제한된 training class를 극복한 일반화 성능을 목표로 함
  - Panoptic segmentation
    - Semantic segmentation
      - Per-pixel class labels
    - Instance segmentation
      - Per-object instance labels
    - Panoptic segmentation
      - Per-pixel class+instance labels
      - [class][instance id]로 label


(a) Image
(b) Semantic Segmentation
(c) Instance Segmentation
(d) Panoptic Segmentation

Background

- **Stable Diffusion**
  - Diffusion-based High-resolution synthesis
    - Autoencoder 구조를 통해 latent embedding을 압축하여 perceptual image compression
    - 기존의 diffusion model보다 high-resolution 이미지 생성
  - Text-to-image model
    - Text를 Encoder τ로 encoding 하여 condition으로 활용
    - 이외에도 semantic map, text, representations, images 등을 condition으로 활용함



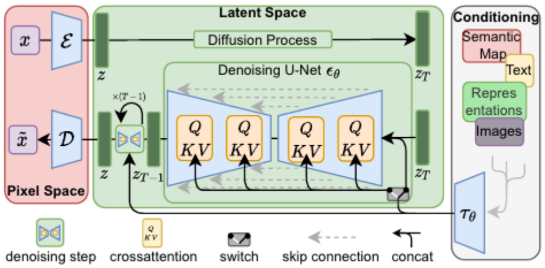


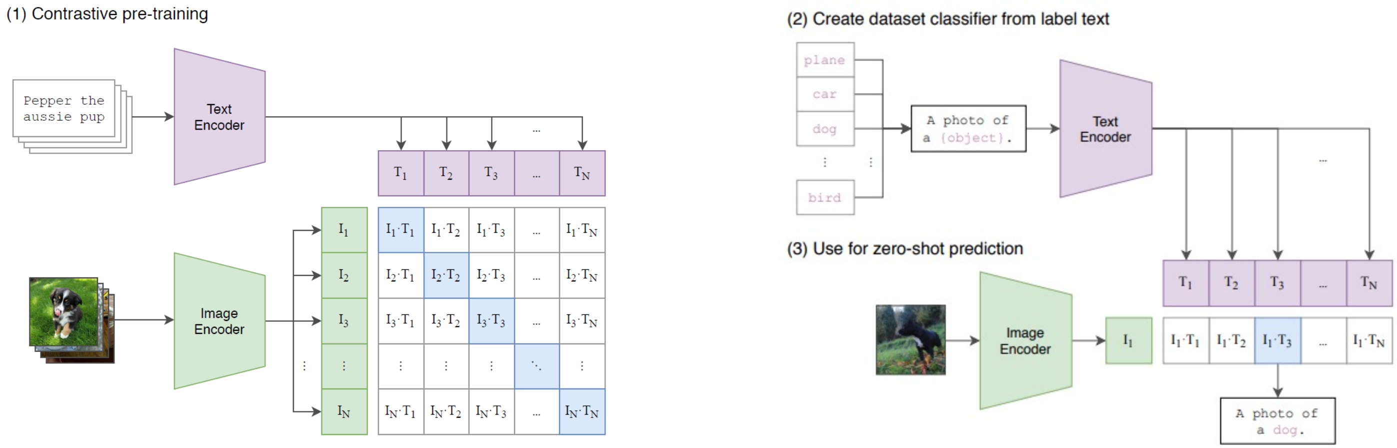
Figure 3. We condition LDMs either via concatenation or by a more general cross-attention mechanism. See Sec. 3.3

Stable diffusion Architecture

- **Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models**
    - CLIP Embedding
    - NVIDIA
    - CVPR 2023 Highlight

CLIP Embedding

- 이미지와 텍스트 사이의 유사도를 추론하는 모델
- Contrastive pre-training
  - Image encoder와 Text encoder를 통과해서 나온 image feature vector와 text feature vector 사이의 연결 관계를 학습
- Zero-shot prediction
  - Image encoder로 image feature를 추출하고, 모든 class label을 text encoder에 통과시켜 text feature를 추출
  - N개의 text feature 중 image feature와 가장 높은 상관관계를 가지는 text를 분류 결과로 선택



(1) Contrastive pre-training

(2) Create dataset classifier from label text

(3) Use for zero-shot prediction

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$

Introduction

- 높은 이해도를 위해 open-vocabulary가 활용되고 있지만, panoptic segmentation에서는 잘 활용되지 못하고 있음
- 본 논문에서는 **large-scale text-image diffusion(Stable Diffusion)**과 **discriminative model(CLIP)**을 활용해 panoptic segmentation을 하는 **ODISE**를 제안

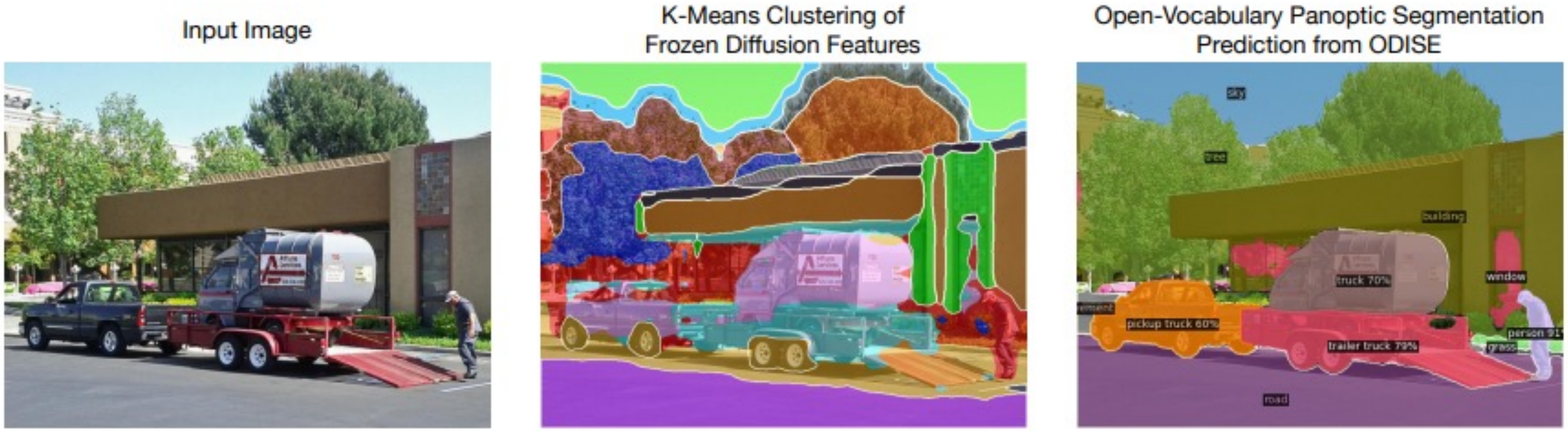

Input Image
K-Means Clustering of Frozen Diffusion Features
Open-Vocabulary Panoptic Segmentation Prediction from ODISE
sky
tree
building
truck 70%
window
pavement
pickup truck 60%
trailer truck 79%
person 91
grass
road

Figure 1. We learn open-vocabulary panoptic segmentation with the internal representation of text-to-image diffusion models. K-Means clustering of the diffusion model's internal representation shows semantically differentiated and localized information wherein objects are well grouped together (middle figure). We leverage these dense and rich diffusion features to perform open-vocabulary panoptic segmentation (right figure).

Open-vocabulary panoptic Segmentation

Method

- Problem definition
  - C_{train}: train 카테고리, C_{test}: test 카테고리
  - C_{test}에는 C_{train}에 없는 카테고리도 포함
  - Test 시에는 이미지별 category label이나 caption이 포함되지 않고, C_{test}만 제공됨

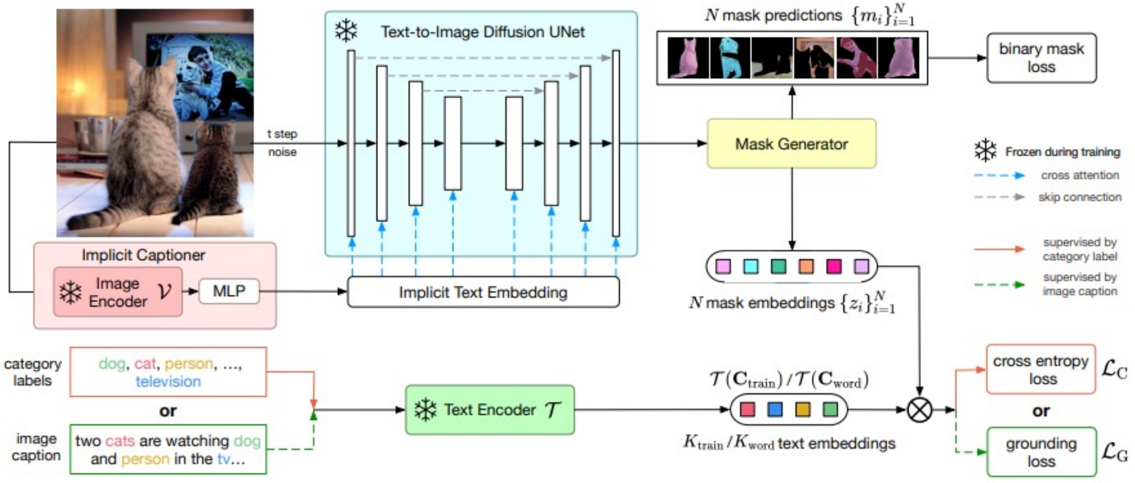

Text-to-Image Diffusion UNet
t step
noise
N mask predictions $\{m_i\}_{i=1}^{N}$
binary mask loss
Mask Generator
Frozen during training
cross attention
skip connection
supervised by category label
supervised by image caption
Implicit Captioner
Image Encoder $\mathcal{V}$
MLP
Implicit Text Embedding
N mask embeddings $\{z_i\}_{i=1}^{N}$
category labels
dog, cat, person, ..., television
or
image caption
two cats are watching dog and person in the tv...
Text Encoder $\mathcal{T}$
$\mathcal{T}(\mathbf{C}_{\text{train}})$ / $\mathcal{T}(\mathbf{C}_{\text{word}})$
K_{train} / K_{word} text embeddings
cross entropy loss
$\mathcal{L}_C$
or
grounding loss
$\mathcal{L}_G$

Figure 2. **ODISE Overview and Training Pipeline**. We first encode the input image into an implicit text embedding with an implicit captioner (image encoder $\mathcal{V}$ and MLP). With the image and its implicit text embedding as input, we extract their diffusion features from a frozen text-to-image diffusion UNet (Sec 3.3). With the UNet's features, a mask generator predicts class-agnostic binary masks and their associated mask embedding features (Sec 3.4). We perform a dot product between the mask embedding features and the text embeddings of training category names (red box) or the nouns of the image's caption (green box) to categorize them. The similarity matrix for mask classification is supervised by either a cross entropy loss with ground truth category labels (red solid path), or via a grounding loss with the paired image captions (green dash path) (Sec 3.5).

ODISE training pipeline

Method

- Text-to-image diffusion model
  - Diffusion-based text-to-image generative model은 denoising process를 위해 UNet 채택
    - UNet은 convolutional blocks, upsampling and downsampling blocks, skip connections, attention blocks로 구성
    - Attention blocks에서 text embedding과 UNet features 간의 cross attention 계산
  - 본 논문에서는 visual representation을 추출하기 위해 diffusion의 single forward pass만 수행
  - Noisy image와 caption이 pair로 들어가면 x에 대한 visual representation f가 caption s에 종속되게 됨

- Implicit Captioner
  - Input image에 내재된 text embedding을 생성하기 위한 network
  - Frozen한 CLIP를 통해 embedding space로 encode 이후, MLP를 통해 implicit text embedding으로 project

$$\begin{aligned} f &= \mathrm{UNet}(x_t, \mathrm{ImplicitCaptioner}(x)) \\ &= \mathrm{UNet}(x_t, \mathrm{MLP} \circ \mathcal{V}(x)). \end{aligned} \tag{3}$$

Implicit Captioner

Method

- Mask Generator
  - Input
    - f
  - Output
    - N class-agnostic binary masks $\{m_i\}_{i=1}^{N}$
    - N mask embeddings $\{z_i\}_{i=1}^{N}$

Method

- Mask classification
  - 예측된 binary mask를 open vocabulary의 category label에 배정하기 위해, text-image discriminative model(CLIP)을 활용
  - 대표적인 두가지 supervision signal
    - Category label supervision
      - Training시 각 마스크의 category label GT가 사용 가능하다고 가정
      - 각 mask embedding feature z_i에 대해 ground truth category y_i를 붙임
      - Mask embedding feature z_i가 K_{train} classes에 속할 확률 계산
      - CrossEntropyLoss 사용

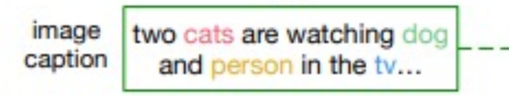
$$\mathcal{L}_{\mathrm{C}} = \frac{1}{N}\sum_{i}^{N} \mathrm{CrossEntropy}(\mathbf{p}(z_i, \mathbf{C}_{\mathrm{train}}), y_i), \qquad (5)$$

$$\mathbf{p}(z_i, \mathbf{C}_{\mathrm{train}}) = \mathrm{Softmax}(z_i \cdot \mathcal{T}(\mathbf{C}_{\mathrm{train}})/\tau), \qquad (6)$$

Category label supervision loss
$\tau(C_{train})$: encoded train categories
τ: learnable parameter

Method

- Mask classification
  - 예측된 binary mask를 open vocabulary의 category label에 배정하기 위해, text-image discriminative model(CLIP)을 활용
  - 대표적인 두가지 supervision signal
    - Image caption supervision
      - Training시 각 마스크의 category label GT가 없다고 가정
      - 대신 image별 caption에서 명사를 추출해서 GT label처럼 활용
      - Grounding loss 사용



$$g(x^{(m)}, s^{(m)}) = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N}\mathbf{p}(z_i, \mathbf{C}_{\text{word}})_k \cdot \langle z_i, \mathcal{T}(w_k)\rangle, \tag{7}$$

The similarity between each image-caption pair

$$\begin{aligned}\mathcal{L}_{\text{G}} = &-\frac{1}{B}\sum_{m=1}^{B}\log\frac{\exp(g(x^{(m)}, s^{(m)})/\tau)}{\sum_{n=1}^{B}\exp(g(x^{(m)}, s^{(n)})/\tau)} \\ &-\frac{1}{B}\sum_{m=1}^{B}\log\frac{\exp(g(x^{(m)}, s^{(m)})/\tau)}{\sum_{n=1}^{B}\exp(g(x^{(n)}, s^{(m)})/\tau)},\end{aligned} \tag{8}$$

Image caption supervision loss

Method

- Open-vocabulary inference
  - Inference시 caption과 label은 쓸 수 없고 test categories C_{test}만 사용 가능
    - Implicit captioner 활용
  - Predicted mask를 test category로 classify하기 위해 text-image discriminative model(CLIP)의 encoder를 활용
    - Image encoder로 image x를 encode해서 feature map으로 변환
    - Mask pooled image feature 계산
    - Mask pooled probabilities와 mask probabilities를 geometric mean하여 계산

$$z_i' = \text{MaskPooling}(\mathcal{V}(x), m_i). \qquad (9)$$

Mask pooled image feature

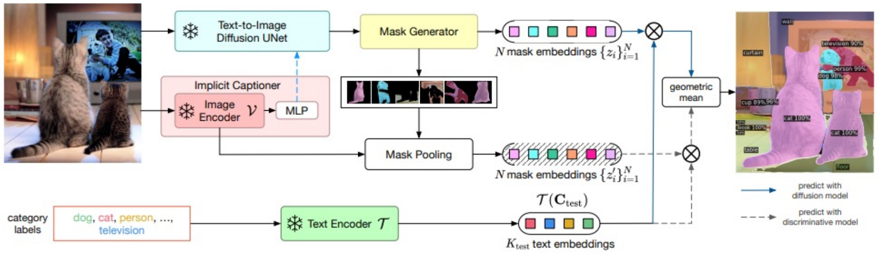
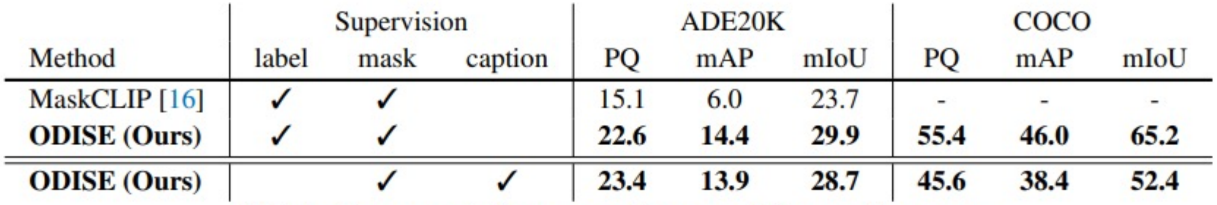

Text-to-Image Diffusion UNet
Mask Generator
N mask embeddings $\{z_i\}_{i=1}^N$
Implicit Captioner
Image Encoder $\mathcal{V}$
MLP
Mask Pooling
N mask embeddings $\{z_i'\}_{i=1}^N$
geometric mean
$\mathcal{T}(\mathbf{C}_{\text{test}})$
category labels
dog, cat, person, ..., television
Text Encoder $\mathcal{T}$
K_{test} text embeddings
wall
curtain
television 90%
person 99%
dog 98%
cup 89% 96%
cat 100%
book 100%
cat 100%
table
floor
predict with diffusion model
predict with discriminative model

Figure 3. **Open-Vocabulary Inference Pipeline**. To classify each mask embedding into testing categories $\mathbf{C}_{\text{test}}$, we compute its similarity with the text encoder $\mathcal{T}$ embedding of category names. Besides the mask embeddings from text-to-image diffusion model $\{z_i\}_{i=1}^N$, we also perform mask pooling on the features of image encoder $\mathcal{V}$ from text-image discriminative model to get $\{z_i'\}_{i=1}^N$. We fuse the prediction of diffusion model (blue solid path) and discriminative model (grey dash path) with geometric mean.

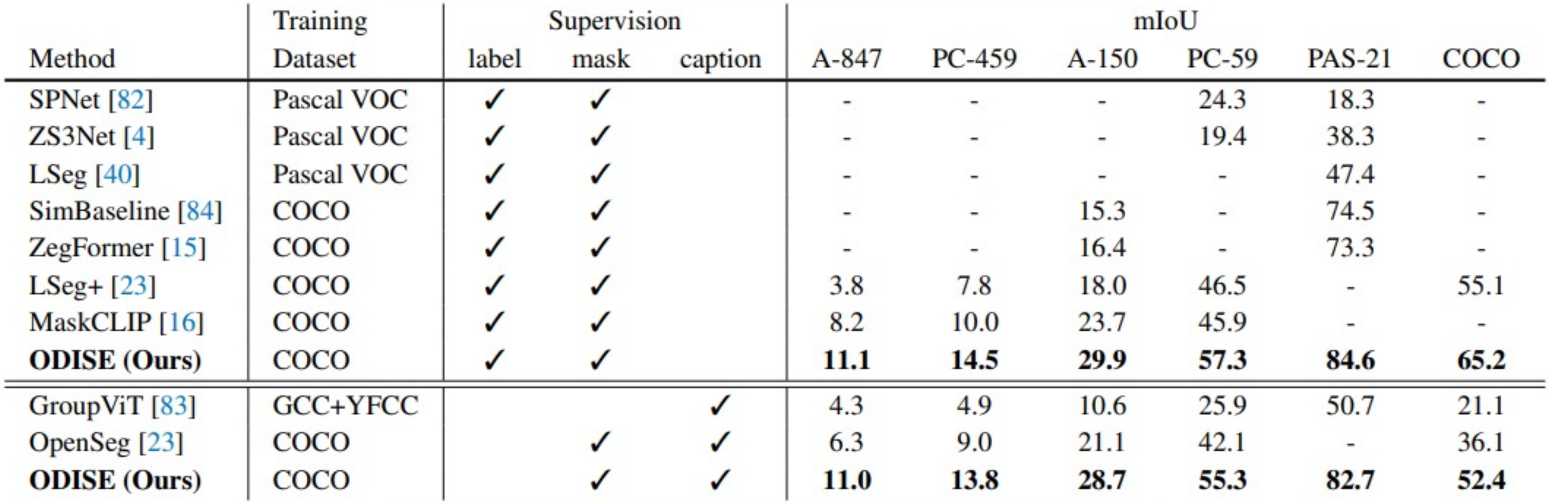
Open-vocabulary inference pipeline

Experiments



Method	Supervision label	Supervision mask	Supervision caption	ADE20K PQ	ADE20K mAP	ADE20K mIoU	COCO PQ	COCO mAP	COCO mIoU
MaskCLIP [16]	✓	✓		15.1	6.0	23.7	-	-	-
ODISE (Ours)	✓	✓		**22.6**	**14.4**	**29.9**	**55.4**	**46.0**	**65.2**
ODISE (Ours)		✓	✓	**23.4**	**13.9**	**28.7**	**45.6**	**38.4**	**52.4**

Table 1. **Open-vocabulary panoptic segmentation performance.**



Method	Training Dataset	Supervision label	Supervision mask	Supervision caption	mIoU A-847	mIoU PC-459	mIoU A-150	mIoU PC-59	mIoU PAS-21	mIoU COCO
SPNet [82]	Pascal VOC	✓	✓		-	-	-	24.3	18.3	-
ZS3Net [4]	Pascal VOC	✓	✓		-	-	-	19.4	38.3	-
LSeg [40]	Pascal VOC	✓	✓		-	-	-	-	47.4	-
SimBaseline [84]	COCO	✓	✓		-	-	15.3	-	74.5	-
ZegFormer [15]	COCO	✓	✓		-	-	16.4	-	73.3	-
LSeg+ [23]	COCO	✓	✓		3.8	7.8	18.0	46.5	-	55.1
MaskCLIP [16]	COCO	✓	✓		8.2	10.0	23.7	45.9	-	-
ODISE (Ours)	COCO	✓	✓		**11.1**	**14.5**	**29.9**	**57.3**	**84.6**	**65.2**
GroupViT [83]	GCC+YFCC			✓	4.3	4.9	10.6	25.9	50.7	21.1
OpenSeg [23]	COCO		✓	✓	6.3	9.0	21.1	42.1	-	36.1
ODISE (Ours)	COCO		✓	✓	**11.0**	**13.8**	**28.7**	**55.3**	**82.7**	**52.4**

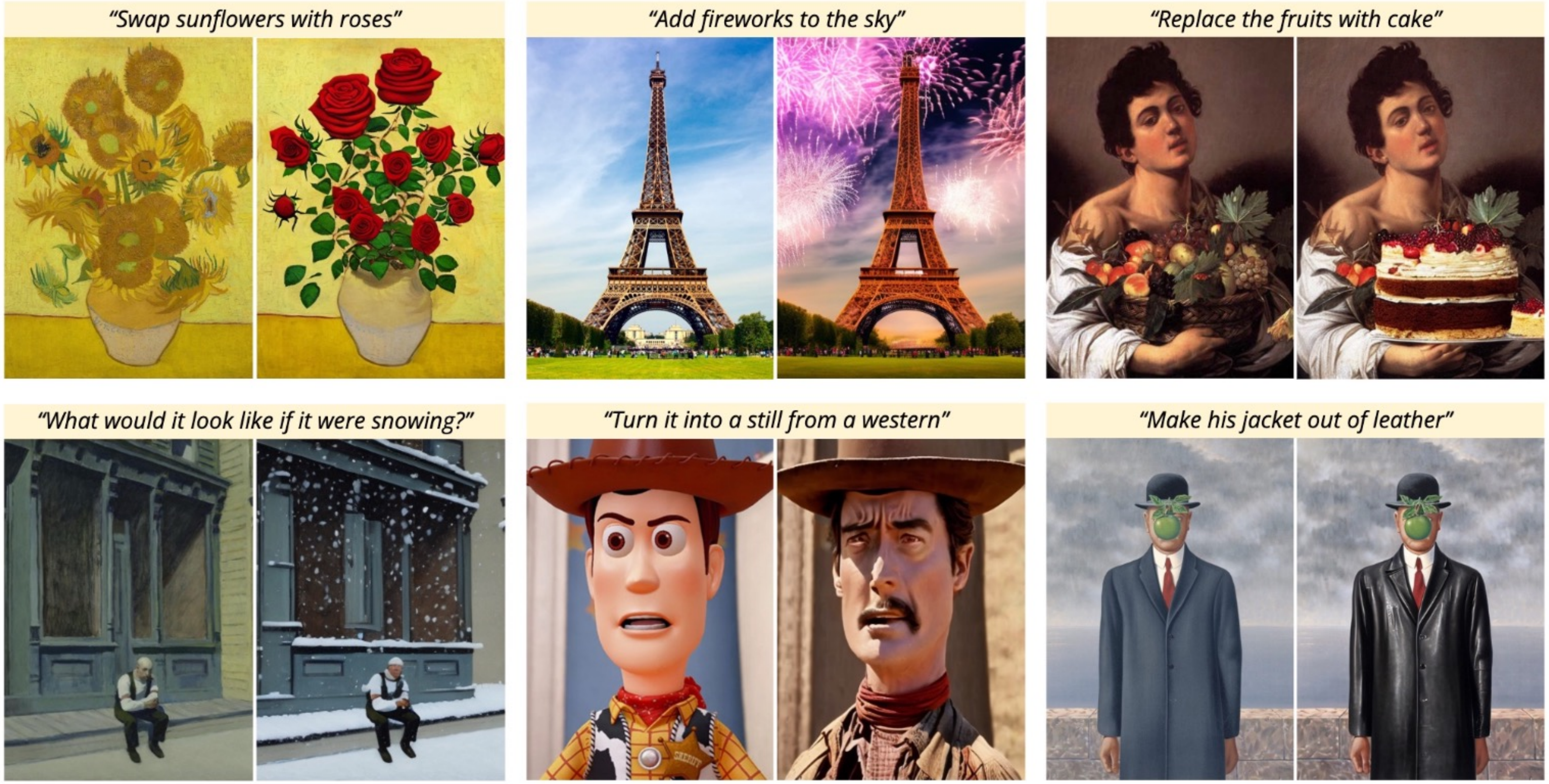
Table 2. **Open-vocabulary semantic segmentation performance.**

Conclusions

- Downstream task에서 frozen internal representation of large-scale text-to-image diffusion models의 가능성을 제안
- Text-to-image diffusion model이 plausible image generation 뿐만 아니라 semantic representation 학습에 뛰어나다는 것을 증명

- **InstructPix2Pix: Learning to Follow Image Editing Instructions**
  - CVPR 2023 Highlight

InstructPix2Pix



"Swap sunflowers with roses"

"Add fireworks to the sky"

"Replace the fruits with cake"

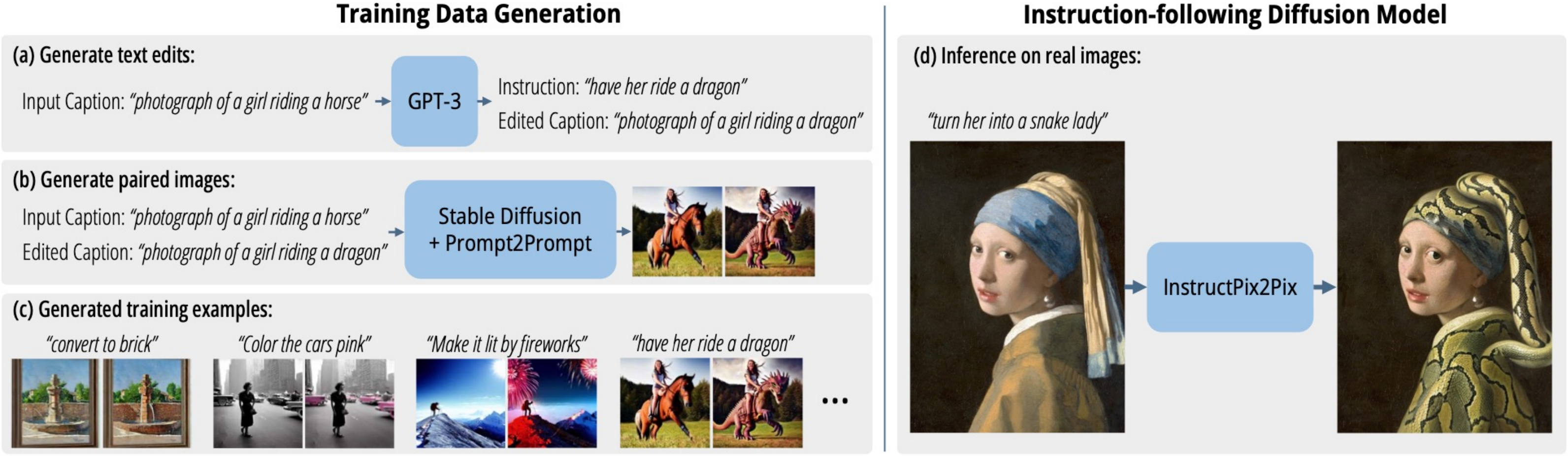
"What would it look like if it were snowing?"

"Turn it into a still from a western"

"Make his jacket out of leather"

Introduction

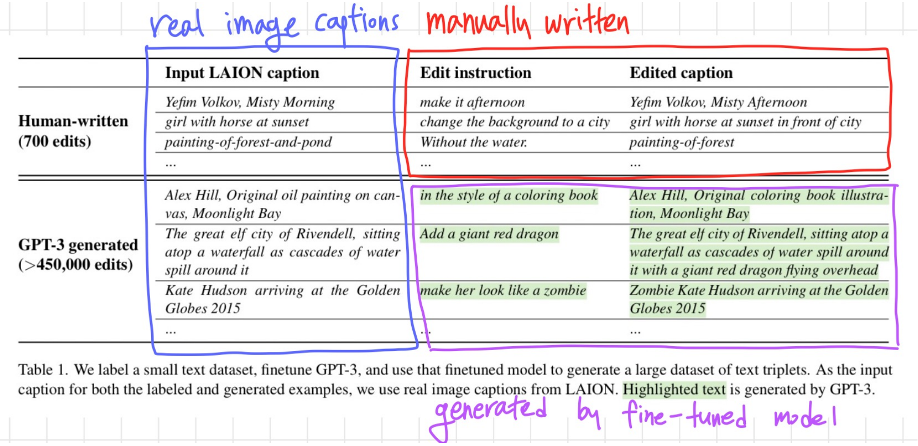
- Human-written instruction 기반 image editing을 위한 학습 방법론을 제시
- 학습 데이터 제작을 위해 GPT-3와 Stable Diffusion 결합
  - Generated data로 학습
  - Real image와 user-written instruction으로 inference


Training Data Generation
(a) Generate text edits:
Input Caption: "photograph of a girl riding a horse"
GPT-3
Instruction: "have her ride a dragon"
Edited Caption: "photograph of a girl riding a dragon"
(b) Generate paired images:
Input Caption: "photograph of a girl riding a horse"
Edited Caption: "photograph of a girl riding a dragon"
Stable Diffusion + Prompt2Prompt
(c) Generated training examples:
"convert to brick"
"Color the cars pink"
"Make it lit by fireworks"
"have her ride a dragon"
...
Instruction-following Diffusion Model
(d) Inference on real images:
"turn her into a snake lady"
InstructPix2Pix

Method

- Instruction-based image editing을 supervised problem으로 정의
  - Instruction과 image의 paired training dataset을 생성
  - Image editing diffusion model을 학습

- Generating instructions and paired captions
  - **Large language model (GPT-3) 활용**
    - Image captions을 input으로, editing instruction과 the resulting text caption을 output으로 생성
    - Small human-written dataset으로 finetuning
      - Finetuning dataset 제작을 위해, 700 개의 input caption에 대한 instruction과 output caption을 manually하게 작성

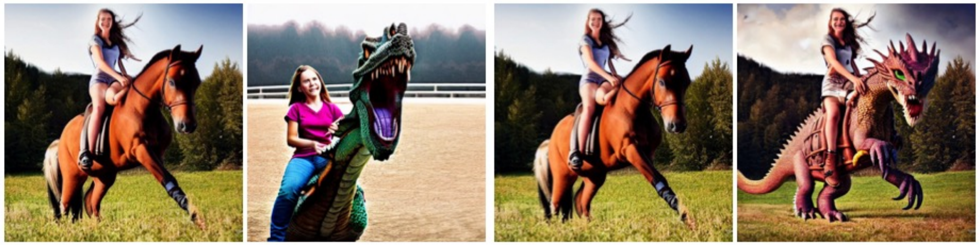

real image captions
manually written
generated by fine-tuned model

	Input LAION caption	Edit instruction	Edited caption
Human-written (700 edits)	*Yefim Volkov, Misty Morning*	*make it afternoon*	*Yefim Volkov, Misty Afternoon*
	girl with horse at sunset	*change the background to a city*	*girl with horse at sunset in front of city*
	painting-of-forest-and-pond	*Without the water.*	*painting-of-forest*
	...	...	...
GPT-3 generated (>450,000 edits)	*Alex Hill, Original oil painting on canvas, Moonlight Bay*	*in the style of a coloring book*	*Alex Hill, Original coloring book illustration, Moonlight Bay*
	The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it	*Add a giant red dragon*	*The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it with a giant red dragon flying overhead*
	Kate Hudson arriving at the Golden Globes 2015	*make her look like a zombie*	*Zombie Kate Hudson arriving at the Golden Globes 2015*
	...	...	...

Table 1. We label a small text dataset, finetune GPT-3, and use that finetuned model to generate a large dataset of text triplets. As the input caption for both the labeled and generated examples, we use real image captions from LAION. Highlighted text is generated by GPT-3.

Method

- Generating paired images from paired captions
  - **Pretrained text-to-image model(Stable Diffusion) 활용**
    - Caption pair(input caption+edited caption)를 image pair로 전환
  - One challenge: text-to-image models는 image consistency를 보장하지 않음
    - 예를 들어, “a picture of a cat”과 “a picture of a black cat " 이 완전히 상이한 이미지 생성
    - Prompt-to-Prompt를 사용하여 해결
      - 몇 개의 denoising steps에서 prompt-to-prompt의 cross attention weight를 활용
  - Caption pair만으로 optimal value를 찾는 것은 쉽지 않음
    - Caption pair마다 100개의 image를 생성한 뒤, CLIP-based metric를 통해 filtering
    - CLIP-based metric: 두 image의 변화와 두 caption의 변화의 일관성을 유지시키는 효과가 있음



(a) Without Prompt-to-Prompt. (b) With Prompt-to-Prompt.

Prompt-to-Prompt를 통한 image consistency

Method

- **InstructPix2Pix**
  - Stable Diffusion 기반 large-scale text-to-image latent diffusion model
  - Image conditioning c_I와 text instruction conditioning c_T가 주어졌을 때, noisy latent z_t에 더해진 noise를 예측하는 네트워크 ϵ_θ를 학습

$$L = \mathbb{E}_{\mathcal{E}(x),\mathcal{E}(c_I),c_T,\epsilon\sim\mathcal{N}(0,1),t}\Big[\|\epsilon - \epsilon_\theta(z_t, t, \mathcal{E}(c_I), c_T))\|_2^2\Big] \quad (1)$$

InstuctPix2Pix objective function

  - Pretrained Stable diffusion checkpoint를 초기화하여 사용
  - 첫번째 convolutional layer에 추가적인 input channels를 더하여 사용
    - z_t와 $\varepsilon(c_I)$를 concat

Method

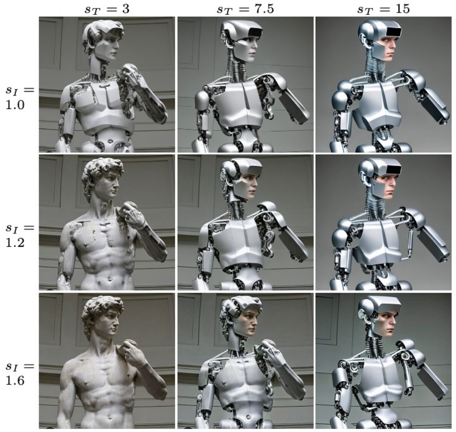
- **Classifier-free guidance for two conditionings**
  - 원래의 Classifier-free: Guidance scale s로 class guidance의 scale을 조정
  - InstructPix2Pix: 두 guidance scale(s_I와 s_T)로 scale을 조정

$$\tilde{e_\theta}(z_t, c) = e_\theta(z_t, \varnothing) + s \cdot (e_\theta(z_t, c) - e_\theta(z_t, \varnothing)) \quad (2)$$

Original classifier-free score estimate

$$\begin{aligned}\tilde{e_\theta}(z_t, c_I, c_T) = {} & e_\theta(z_t, \varnothing, \varnothing) \\ & + s_I \cdot (e_\theta(z_t, c_I, \varnothing) - e_\theta(z_t, \varnothing, \varnothing)) \\ & + s_T \cdot (e_\theta(z_t, c_I, c_T) - e_\theta(z_t, c_I, \varnothing))\end{aligned} \quad (3)$$

InstructPix2Pix score estimate



Edit instruction: *"Turn him into a cyborg!"*

Two conditioning inputs

Method

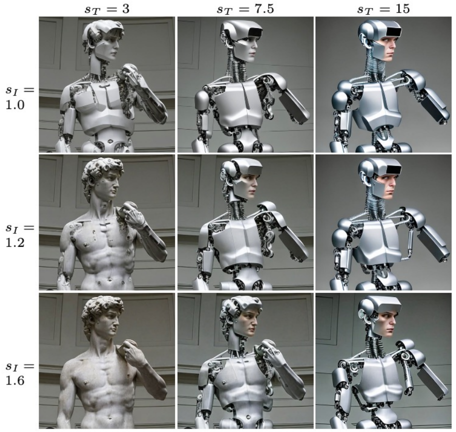
- **Classifier-free guidance for two conditionings**
  - 원래의 Classifier-free: Guidance scale s로 class guidance의 scale을 조정
  - InstructPix2Pix: 두 guidance scale(s_I와 s_T)로 scale을 조정

$$\tilde{e_\theta}(z_t, c) = e_\theta(z_t, \varnothing) + s \cdot (e_\theta(z_t, c) - e_\theta(z_t, \varnothing)) \quad (2)$$

Original classifier-free score estimate

$$\begin{aligned}\tilde{e_\theta}(z_t, c_I, c_T) = {} & e_\theta(z_t, \varnothing, \varnothing) \\ & + s_I \cdot (e_\theta(z_t, c_I, \varnothing) - e_\theta(z_t, \varnothing, \varnothing)) \\ & + s_T \cdot (e_\theta(z_t, c_I, c_T) - e_\theta(z_t, c_I, \varnothing)) \end{aligned} \quad (3)$$

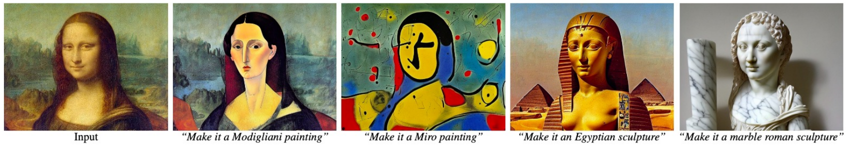
InstructPix2Pix score estimate



Edit instruction: *"Turn him into a cyborg!"*

Two conditioning inputs

Results



Input

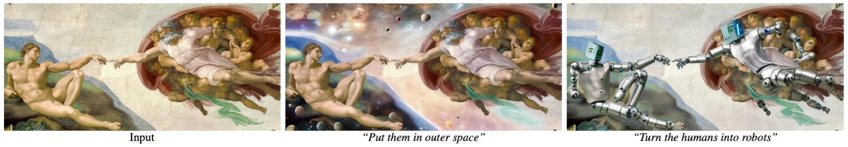
"Make it a Modigliani painting"

"Make it a Miro painting"

"Make it an Egyptian sculpture"

"Make it a marble roman sculpture"

Figure 5. *Mona Lisa* transformed into various artistic mediums.

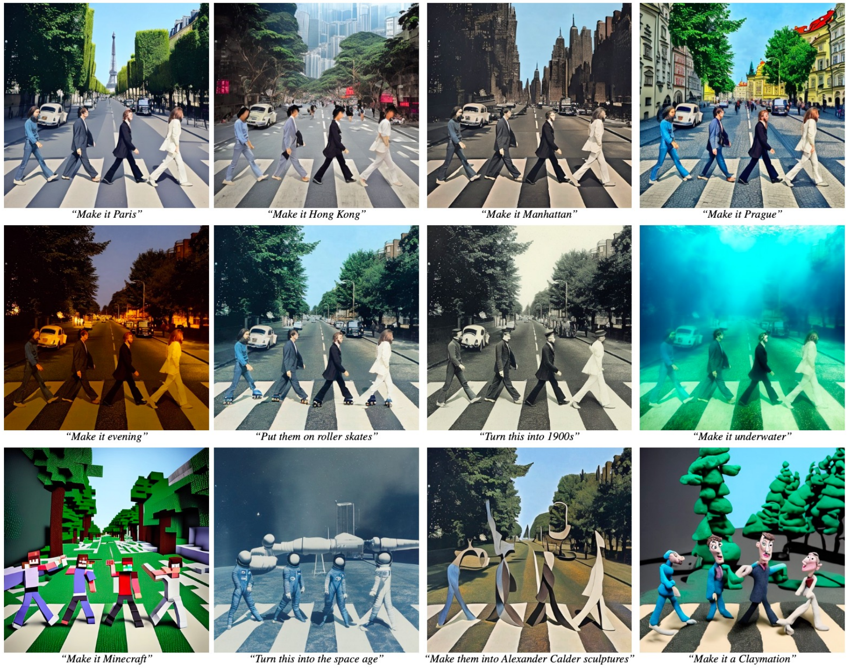


Input

"Put them in outer space"

"Turn the humans into robots"

Figure 6. *The Creation of Adam* with new context and subjects (generated at 768 resolution).



"Make it Paris"

"Make it Hong Kong"

"Make it Manhattan"

"Make it Prague"

"Make it evening"

"Put them on roller skates"

"Turn this into 1900s"

"Make it underwater"

"Make it Minecraft"

"Turn this into the space age"

"Make them into Alexander Calder sculptures"

"Make it a Claymation"

Figure 7. The iconic Beatles *Abbey Road* album cover transformed in a variety of ways.

References

- Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.
- https://www.timothybrooks.com/instruct-pix2pix
- Brooks, Tim, Aleksander Holynski, and Alexei A. Efros. "Instructpix2pix: Learning to follow image editing instructions." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
- Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

감사합니다.