2022 하계 세미나

How to Use VLP Models

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

Presented by
조유빈

Outline

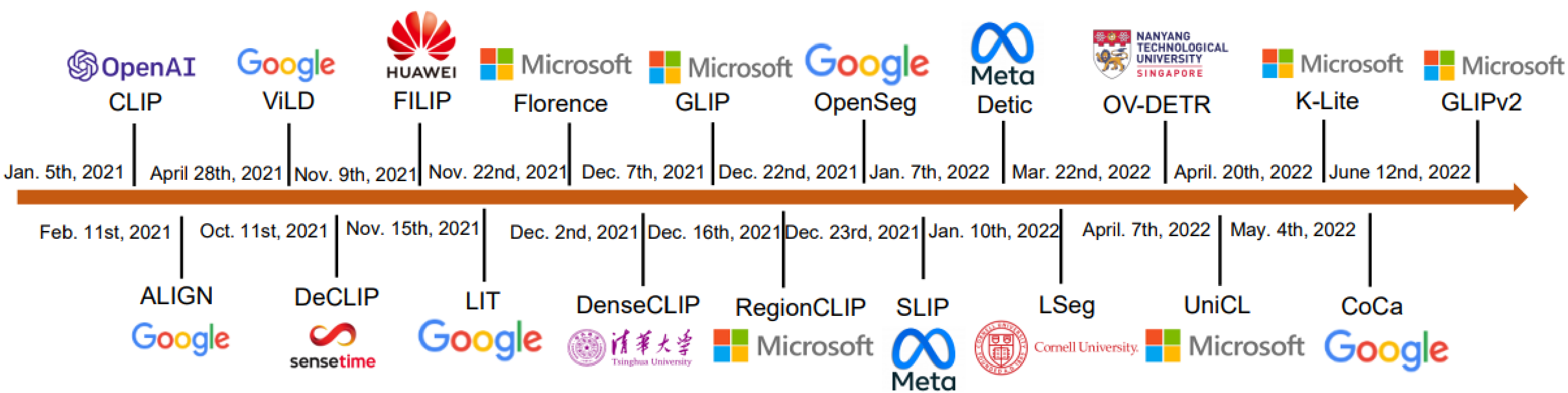
- Background
  - Vision-Language Pretraining (VLP)
  - Contrastive Language-Image Pretraining (CLIP)

- How to use CLIP
  - CLIP-Driven Segmentation
    - CRIS: CLIP-Driven Referring Image Segmentation (CVPR 2022)
  - CLIP-Driven Manipulation
    - CLIP-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields (CVPR 2022)

- Conclusion

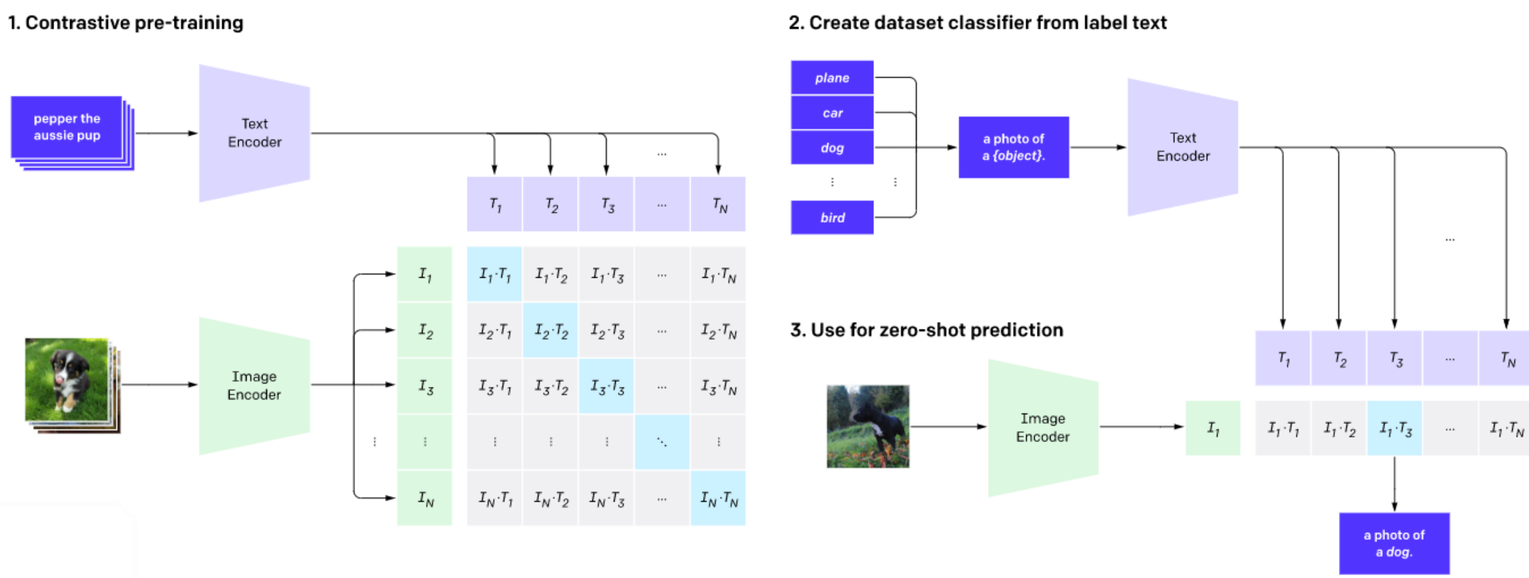
Background

- Vision-Language Pretraining (VLP) Models
  - 대용량의 image-text pair dataset으로 두 모달의 representations를 학습시킨 모델
  - 주로 contrastive learning을 사용하여 학습
    - Positive pair는 유사도가 커지도록, negative pair는 유사도가 작아지도록 학습
    - Supervised / self-supervised / generative learning과 결합하여 학습되기도 함
  - Uni-modal task 또는 vision-language multi-modal task에서 높은 성능을 보임
    - Classification / object detection / captioning / retrieval etc.



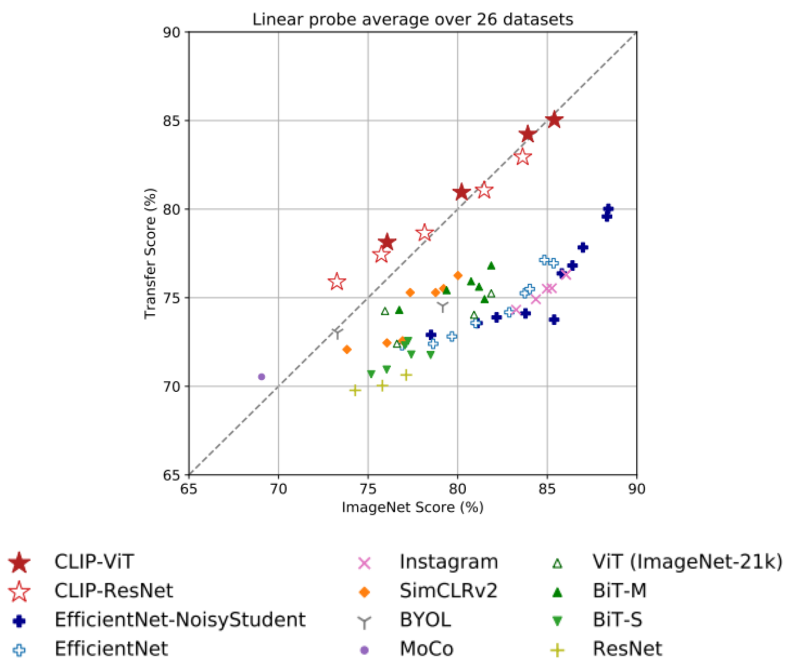
Background

- [1] Contrastive Language-Image Pretraining (CLIP)
  - Contrastive pre-training
    - Aligning two modalities representations in a multi-modal embedding space
      - Maximize the cosine similarity of the image-text embeddings of the N real pairs
      - Minimizing the cosine similarity of the embeddings of the N^2-N incorrect pairs
  - Use for prediction (Image classification)



Background

- [1] Contrastive Language-Image Pretraining (CLIP)
  - Results
    - CLIP’s features are more robust to task shift when compared to models pre-trained on ImageNet
    - Zero shot CLIP model is compared with ResNet-101 that has the same performance on ImageNet validation set



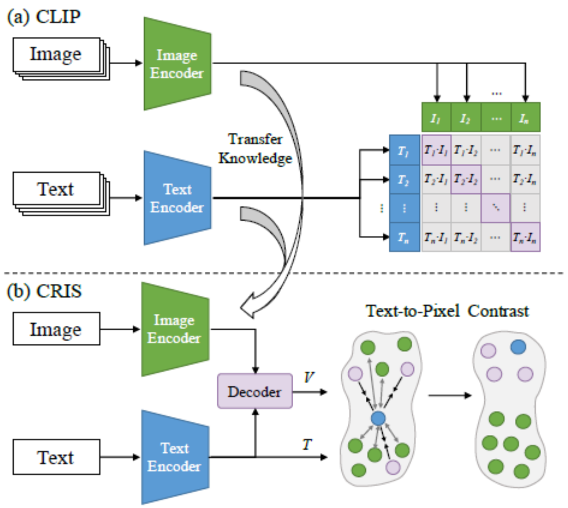
Background

- Using [1] CLIP (Accepted to CVPR 2022)

CLIMS: Cross Language Image Matching for Weakly Supervised Semantic Segmentation
Conditional Prompt Learning for Vision-Language Models
RegionCLIP: Region-based Language-Image Pretraining
DenseCLIP: Language-Guided Dense Prediction with Context-Aware Prompting
CLIP-Event: Connecting Text and Images with Event Structures
ProposalCLIP: Unsupervised Open-Category Object Proposal Generation via Exploiting CLIP Cues
HairCLIP: Design Your Hair by Text and Reference Image
Simple but Effective: CLIP Embeddings for Embodied AI
PointCLIP: Point Cloud Understanding by CLIP
CRIS: CLIP-Driven Referring Image Segmentation
CLIP-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields
CLIP-Forge: Towards Zero-Shot Text-to-Shape Generation
Disentangling visual and written concepts in CLIP
Causal CLIP Fine-tuning for Fashion Product Retrieval
DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation
CLIPstyler: Image Style Transfer with a Single Text Condition
Image Segmentation Using Text and Image Prompts

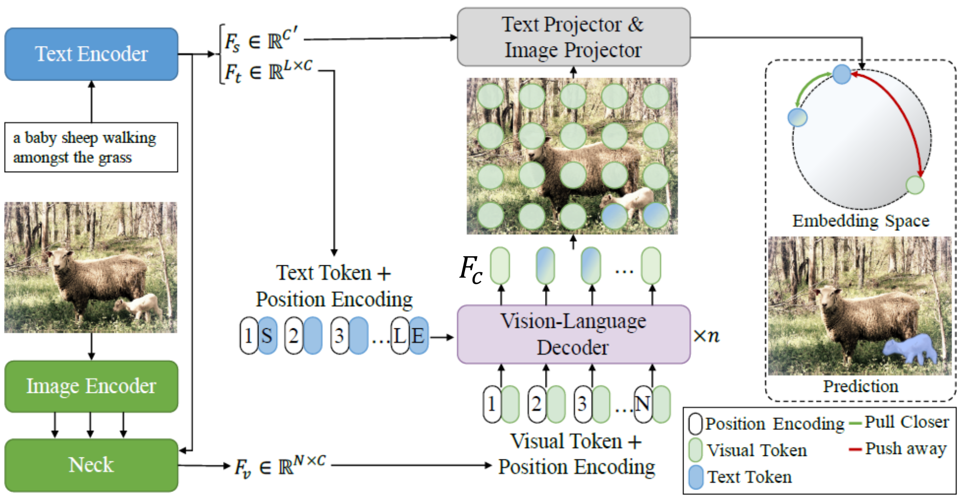
How to Use CLIP

- CLIP-Driven Referring Segmentation : [1] CRIS
  - Referring image segmentation
    - Not limited to indicating specific categories but finding a particular region according to the input language expression
  - CLIP model learns powerful image-level visual concepts by aligning the textual representation with the image-level representation
  - Transfer the knowledge of the CLIP model from image level to pixel level
    - Visual-language decoder
    - Text-to-pixel contrastive learning



How to Use CLIP

- CLIP-Driven Referring Segmentation : [1] CRIS
  - Cross-modal Neck
    - (Multiple visual features, global textual representation F_s) → multi-modal features
    - (Multi-modal features , 2D spatial coordinate feature) → pixel-level visual features F_v
  - Vision-language decoder
    - Propagate fine-grained semantic information from textual features to pixel-level visual features
    - Multi-head self-attention (MHSA) : capture global contextual information
    - Multi-head cross-attention (MHCA) : propagate fine-grained semantic information into F'_v



$$F'_v = MHSA(LN(F_v)) + F_v$$

$$MHSA(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

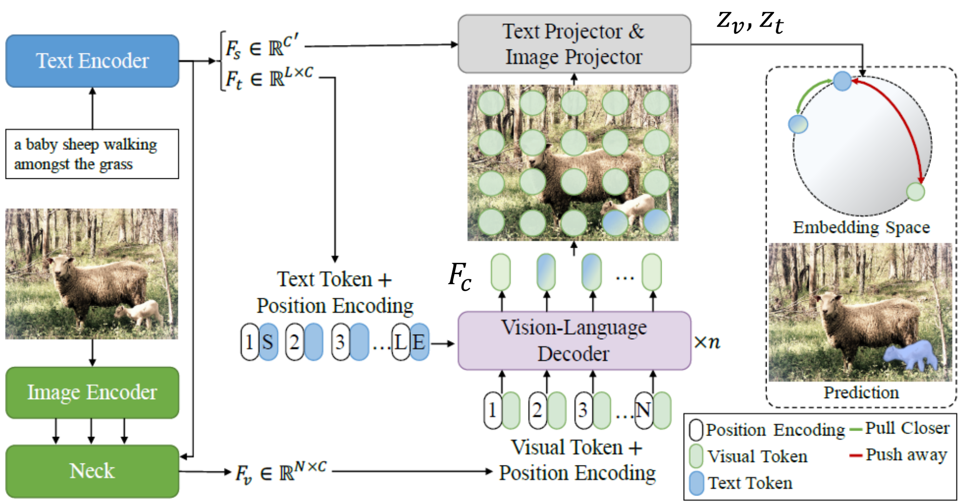
$$F'_c = MHCA(LN(F'_v), F_t) + F'_v$$

$$F_c = MLP(LN(F'_c)) + F'_c$$

F_c : evolved multi-modal feature

How to Use CLIP

- CLIP-Driven Referring Segmentation : [1] CRIS
  - Text-to-pixel contrastive learning
    - Two projectors transform F_c and F_s into the same feature dimension → $z_v^{N\times D}$, z_t^{D}
    - Use text-to-pixel contrastive loss
      - Align text features and the corresponding pixel-level features
      - Distinguish irrelevant pixel-level features in the multi-modal embedding space



$$L^{i}_{con}(z_t, z^{i}_v) = \begin{cases} -\log(\sigma(z_t \cdot z^{i}_v)) & i \in P \\ -\log(1-\sigma(z_t \cdot z^{i}_v)) & i \in N \end{cases}$$

$$L_{con}(z_t, z_v) = \frac{1}{|P \cup N|} \sum_{i \in P \cup N} L^{i}_{con}(z_t, z^{i}_v)$$

Final segmentation results $= \sigma(z_t \cdot z_v)$

How to Use CLIP

- CLIP-Driven Referring Segmentation : [1] CRIS
  - Results

Dataset	*Con.*	*Dec.*	*n*	IoU	Pr@50	Pr@60	Pr@70	Pr@80	Pr@90
RefCOCO	-	-	-	62.66	72.55	67.29	59.53	43.52	12.72
	✓	-	-	64.64	74.89	69.58	61.70	45.50	13.31
	-	✓	1	66.31	77.66	72.99	65.67	48.43	14.81
	✓	✓	1	68.66	80.16	75.72	68.82	51.98	15.94
	✓	✓	2	69.13	80.96	76.60	69.67	52.23	16.09
	✓	✓	3	**69.52**	**81.35**	**77.54**	**70.79**	**52.65**	16.21
	✓	✓	4	69.18	80.99	76.74	69.32	52.57	**16.37**
RefCOCO+	-	-	-	50.17	54.55	47.69	40.19	28.75	8.21
	✓	-	-	53.15	58.28	53.74	46.67	34.01	9.30
	-	✓	1	54.73	63.31	58.89	52.46	38.53	11.70
	✓	✓	1	59.97	69.19	64.85	58.17	43.47	13.39
	✓	✓	2	60.75	70.69	66.83	60.74	45.69	13.42
	✓	✓	3	**61.39**	**71.46**	**67.82**	**61.80**	**47.00**	**15.02**
	✓	✓	4	61.15	71.05	66.94	61.25	46.98	14.97
G-Ref	-	-	-	49.24	53.33	45.49	36.58	23.90	6.92
	✓	-	-	52.67	59.27	52.45	44.12	29.53	8.80
	-	✓	1	51.46	58.68	53.33	45.61	31.78	10.23
	✓	✓	1	57.82	66.28	60.99	53.21	38.58	13.38
	✓	✓	2	58.40	67.30	61.72	54.70	39.67	13.40
	✓	✓	3	**59.35**	**68.93**	**63.66**	**55.45**	**40.67**	**14.40**
	✓	✓	4	58.79	67.91	63.11	55.43	39.81	13.48

< Ablation studies >

Method	Backbone	RefCOCO			RefCOCO+			G-Ref	
		val	test A	test B	val	test A	test B	val	test
RMI* [25]	ResNet-101	45.18	45.69	45.57	29.86	30.48	29.50	-	-
DMN [33]	ResNet-101	49.78	54.83	45.13	38.88	44.22	32.29	-	-
RRN* [22]	ResNet-101	55.33	57.26	53.95	39.75	42.15	36.11	-	-
MAttNet [50]	ResNet-101	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61
NMTree [26]	ResNet-101	56.59	63.02	52.06	47.40	53.01	41.56	46.59	47.88
CMSA* [49]	ResNet-101	58.32	60.61	55.09	43.76	47.60	37.89	-	-
Lang2Seg [5]	ResNet-101	58.90	61.77	53.81	-	-	-	46.37	46.95
BCAN* [16]	ResNet-101	61.35	63.37	59.57	48.57	52.87	42.13	-	-
CMPC* [17]	ResNet-101	61.36	64.53	59.64	49.56	53.44	43.23	-	-
LSCM* [18]	ResNet-101	61.47	64.99	59.55	49.34	53.12	43.50	-	-
MCN [30]	DarkNet-53	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
CGAN [29]	DarkNet-53	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69
EFNet [8]	ResNet-101	62.76	65.69	59.67	51.50	55.24	43.01	-	-
LTS [19]	DarkNet-53	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25
VLT [6]	DarkNet-53	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
CRIS (Ours)	ResNet-50	69.52	72.72	64.70	61.39	67.10	52.48	59.35	59.39
CRIS (Ours)	ResNet-101	**70.47**	**73.18**	**66.10**	**62.27**	**68.08**	**53.68**	**59.87**	**60.36**

< Experiments >

How to Use CLIP

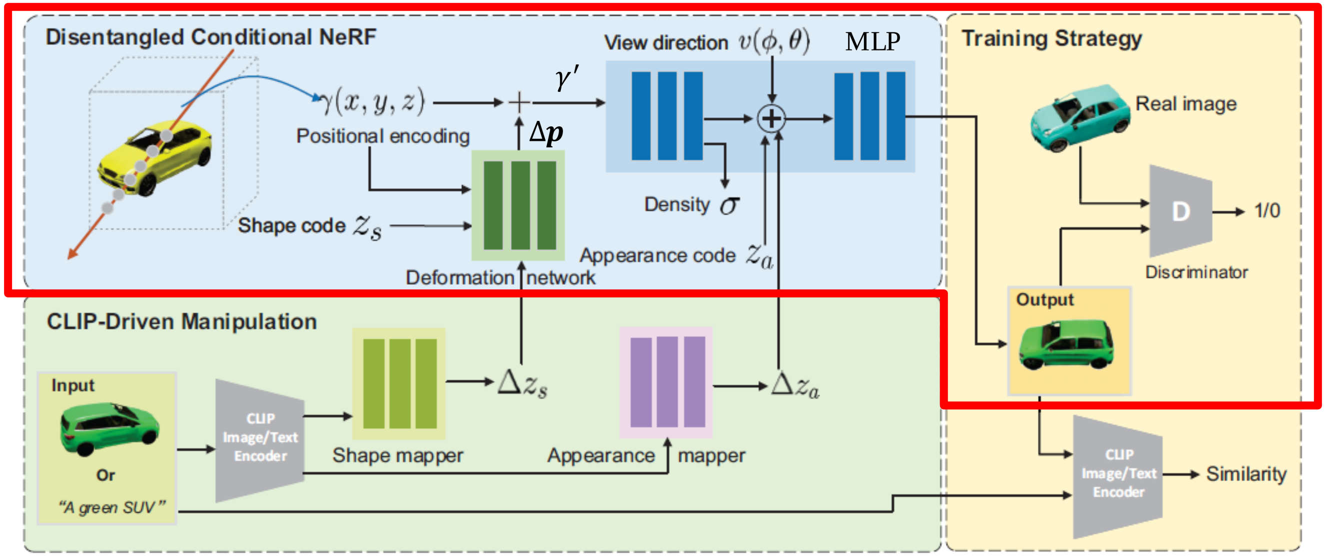
- CLIP-Driven Referring Segmentation : [1] CRIS
  - Results

How to Use CLIP

- CLIP-Driven Manipulation : [1] CLIP-NeRF
  - Conditional NeRF
    - Generative model for a particular object category
    - Conditioned on the latent vectors that dedicatedly control shape and appearance
    - Suffer from mutual intervention between shape and appearance conditions
  - Disentangled conditional NeRF
    - Individual control over both shape and appearance
      - Using CLIP similarity loss for shape mapper and appearance mapper

How to Use CLIP

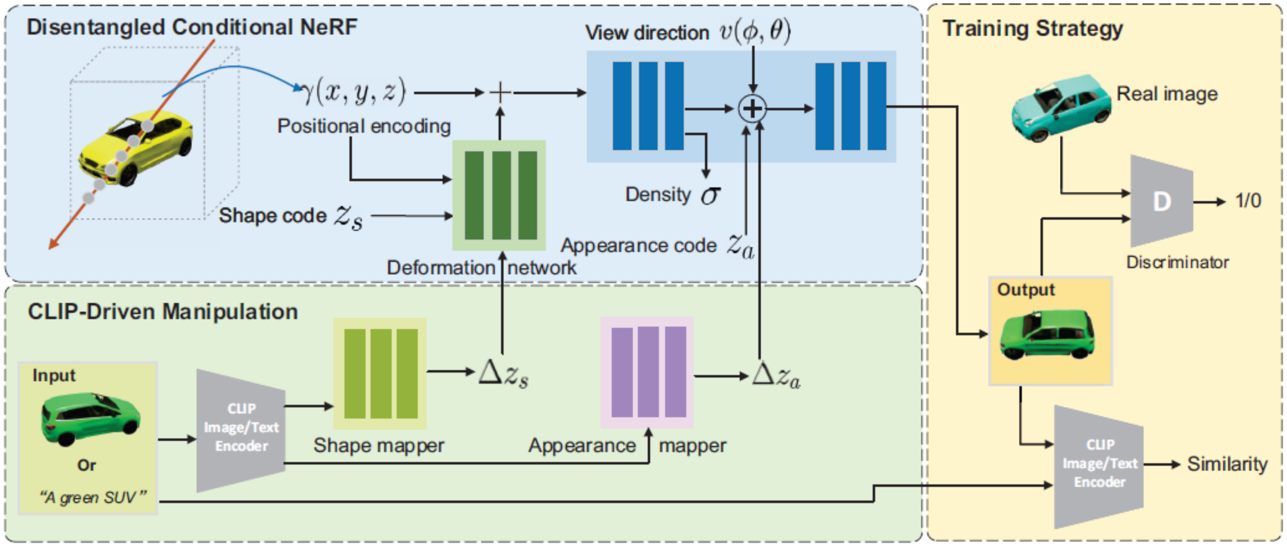
- CLIP-Driven Manipulation : [1] CLIP-NeRF
  - Training disentangled conditional NeRF
    - Inputs : position, view direction, shape code, appearance code
    - Outputs : density, color
    - Shape deformation network : (position $\boldsymbol{p}(x, y, z)$, shape code z_s) → displacement vectors $\Delta\boldsymbol{p}$
    - MLP network
      - ※ manipulating the appearance without touching the shape information (density)



How to Use CLIP

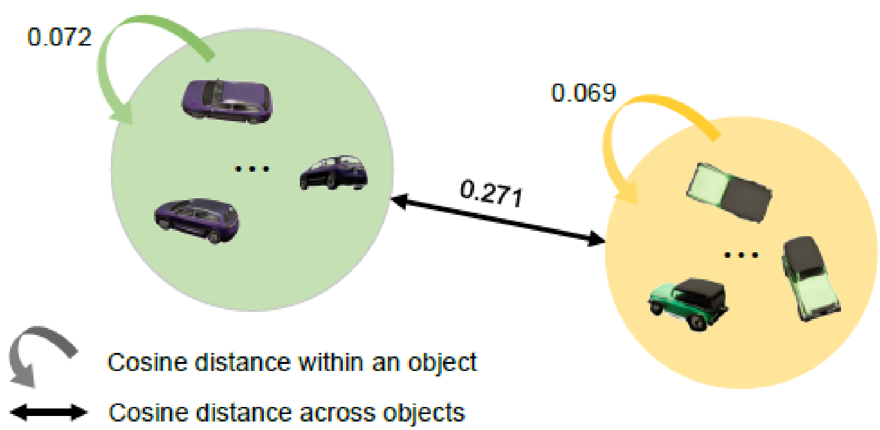
- CLIP-Driven Manipulation : [1] CLIP-NeRF
  - Training mappers
    - Take a feed-forward approach to directly update the condition displacement vectors from the input condition
    - Freeze : generator, discriminator, CLIP encoder
    - Maximize the embedding similarity between a rendered image patch and the input condition

$$z_s' = M_s(\hat{\varepsilon}_t(t)) + z_s, \quad z_a' = M_a(\hat{\varepsilon}_t(t)) + z_a \qquad Loss_{CLIP} = 1 - < \hat{\varepsilon}_i(I), \hat{\varepsilon}_t(t) >$$



How to Use CLIP

- CLIP-Driven Manipulation : [1] CLIP-NeRF
  - CLIP model can support view-consistency representations for 3D-aware applications
    - More sensitive to small object difference than large view variations
      - CLIP feature is stable across different viewpoints
        - ✓ Different views for a same object have higher similarity (small distance)
      - CLIP can distinguish object differences
        - ✓ Different objects have lower similarity (large distance) even in an identical view



How to Use CLIP

- CLIP-Driven Manipulation : [1] CLIP-NeRF
  - Ablation study for disentanglement

How to Use CLIP

- CLIP-Driven Manipulation : [1] CLIP-NeRF
  - Editing results

Source

Orange flowers

Yellow flowers

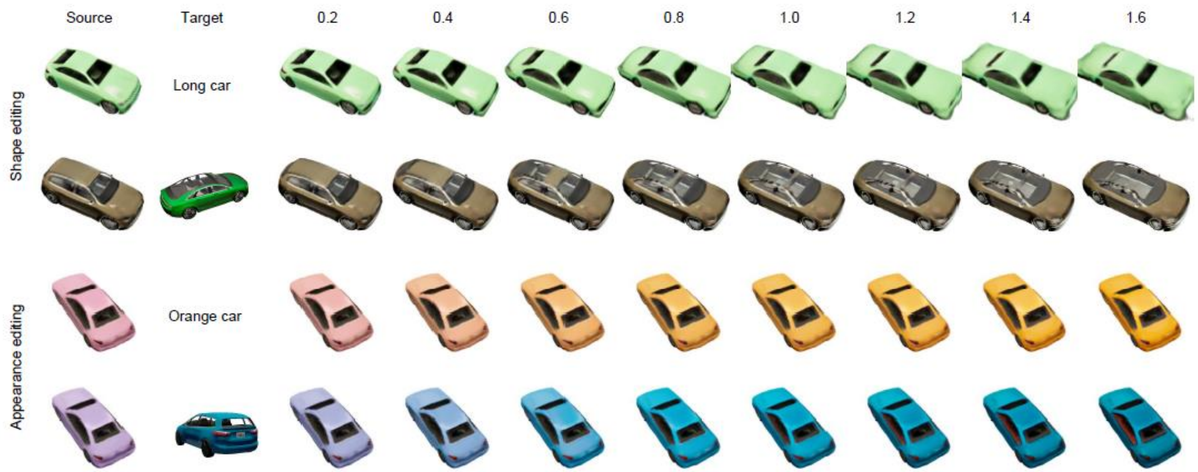
Yellow flowers
with green leaves

How to Use CLIP

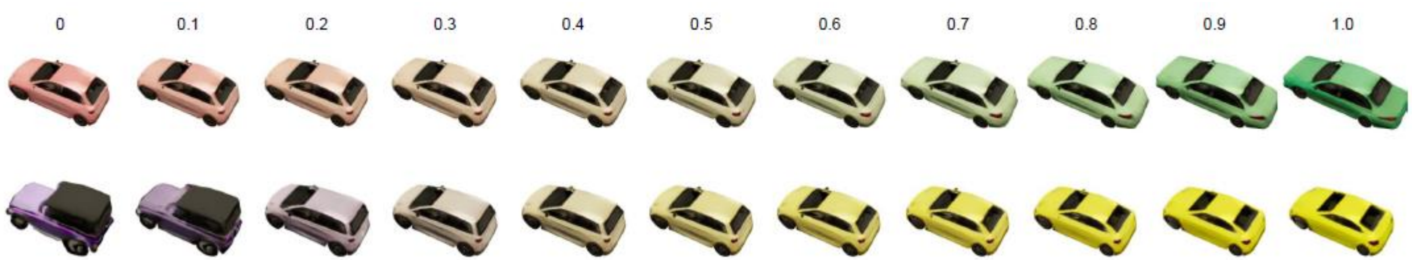
- CLIP-Driven Manipulation : [1] CLIP-NeRF
  - Scaling along editing direction

$$z_s = s \times \Delta z_s + z_s'$$

$$z_a = s \times \Delta z_a + z_a'$$



  - Interpolation $z_{inter} = z^2 \times r + z^1 \times (1 - r)$



Conclusion

- CLIP-Driven Referring Segmentation : [1] CRIS
  - Text-to-pixel contrastive learning
    - Enforce the text feature similar to the related pixel-level features and dissimilar to the irrelevances
- CLIP-Driven Manipulation : [2] CLIP-NeRF
  - Design two code mappers that take a CLIP embedding as input and update the latent codes to reflect the targeted editing
    - Trained with a CLIP-based matching loss to ensure the manipulation accuracy