JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Leveraging Hidden Positives for Unsupervised Semantic Segmentation

Hyun Seok Seong, WonJun Moon, SuBeen Lee, and Jae-Pil Heo
Sungkyunkwan University

Paper ID 6933
THU-AM-291

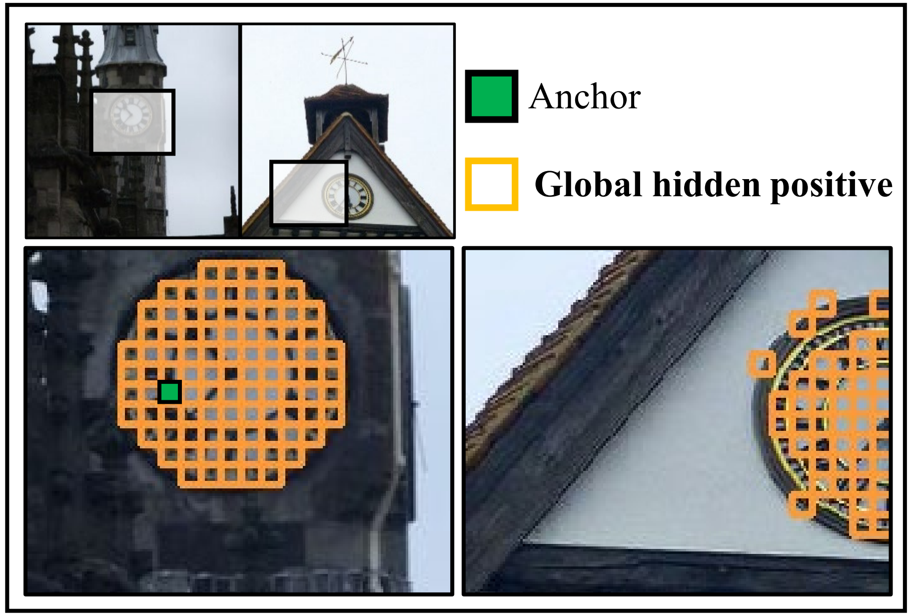
Preview

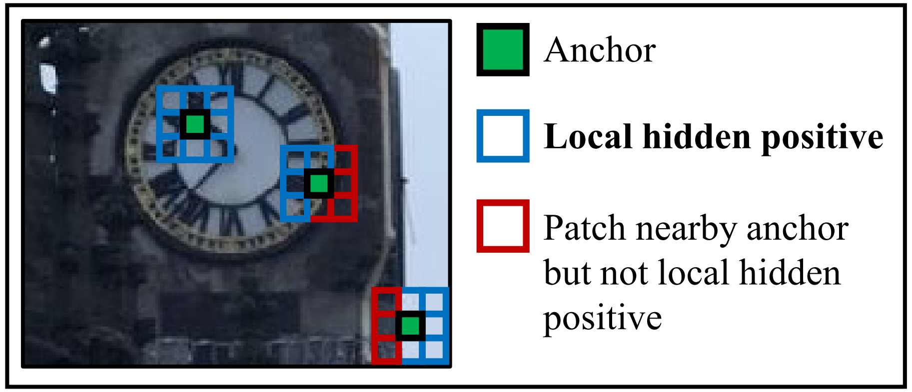
Unsupervised Semantic Segmentation

Without labels

Group 1 Group 2 •••

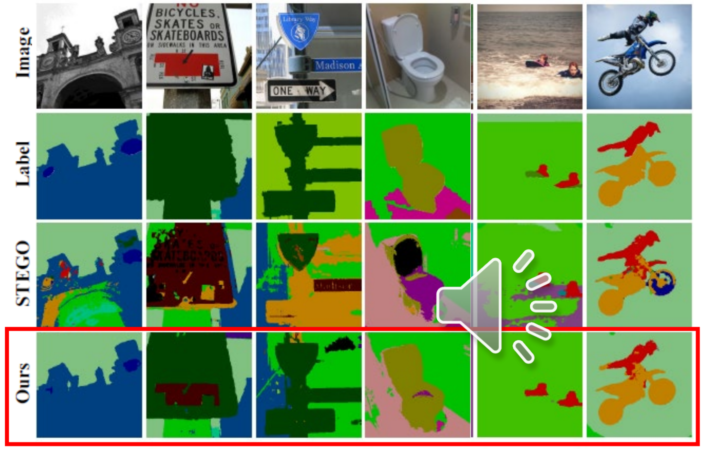
Contrastive learning by discovering hidden positives





Experiment results

Method	Backbone	Unsupervised Acc.	Unsupervised mIoU	Linear Acc.	Linear mIoU
DC [2]	R18+FPN	19.9	-	-	-
MDC [2]	R18+FPN	32.2	9.8	48.6	13.3
IIC [19]	R18+FPN	21.8	6.7	44.5	8.4
PiCIE [8]	R18+FPN	48.1	13.8	54.2	13.9
PiCIE+H [8]	R18+FPN	50.0	14.4	54.8	14.8
DINO	ViT-S/8	28.7	11.3	68.6	33.9
+ TransFGU [37]	ViT-S/8	52.7	17.5	-	-
+ STEGO [14]	ViT-S/8	48.3	24.5	74.4	38.3
+ Ours	ViT-S/8	**57.2**	**24.6**	**75.6**	**42.7**
DINO	ViT-S/16	22.0	8.0	50.3	18.1
+ STEGO [14]	ViT-S/16	52.5	23.7	70.6	34.5
+ Ours	ViT-S/16	**54.5**	**24.3**	**74.1**	**39.1**
SelfPatch	ViT-S/16	35.1	12.3	64.4	28.5
+ STEGO [14]	ViT-S/16	52.4	22.2	72.2	36.0
+ Ours	ViT-S/16	**56.1**	**23.2**	**74.9**	**41.3**



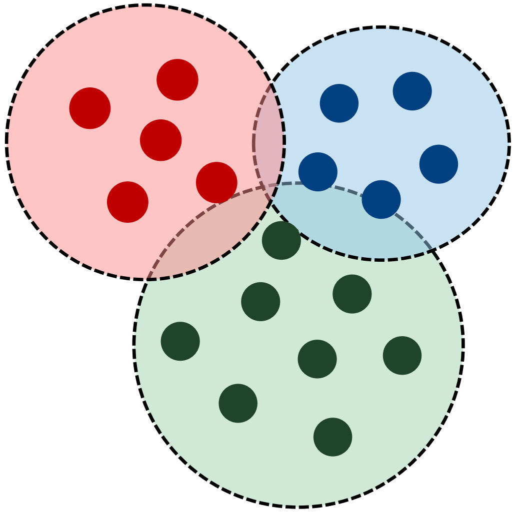
Unsupervised Semantic Segmentation (USS)

- Capturing pixel-level semantics from unlabeled data.

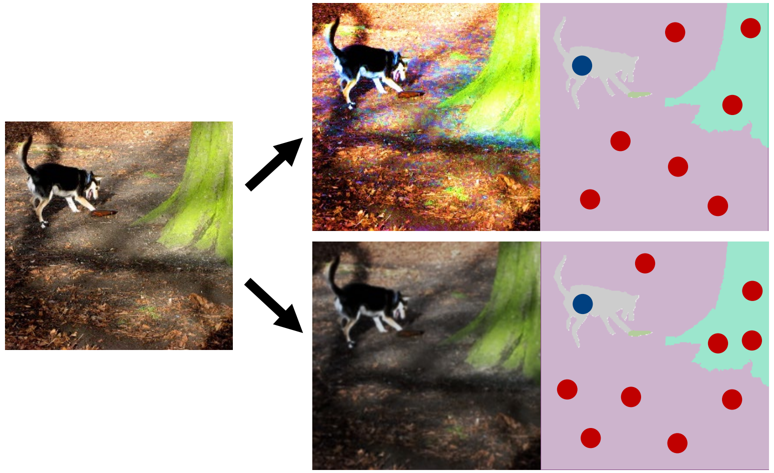
Group 1 Group 2 Group 3 Group 4 •••

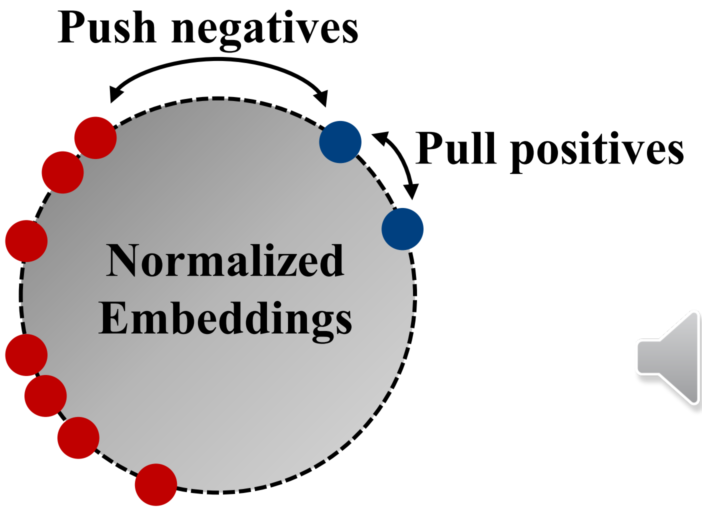
Common approaches

Clustering



Contrastive Learning

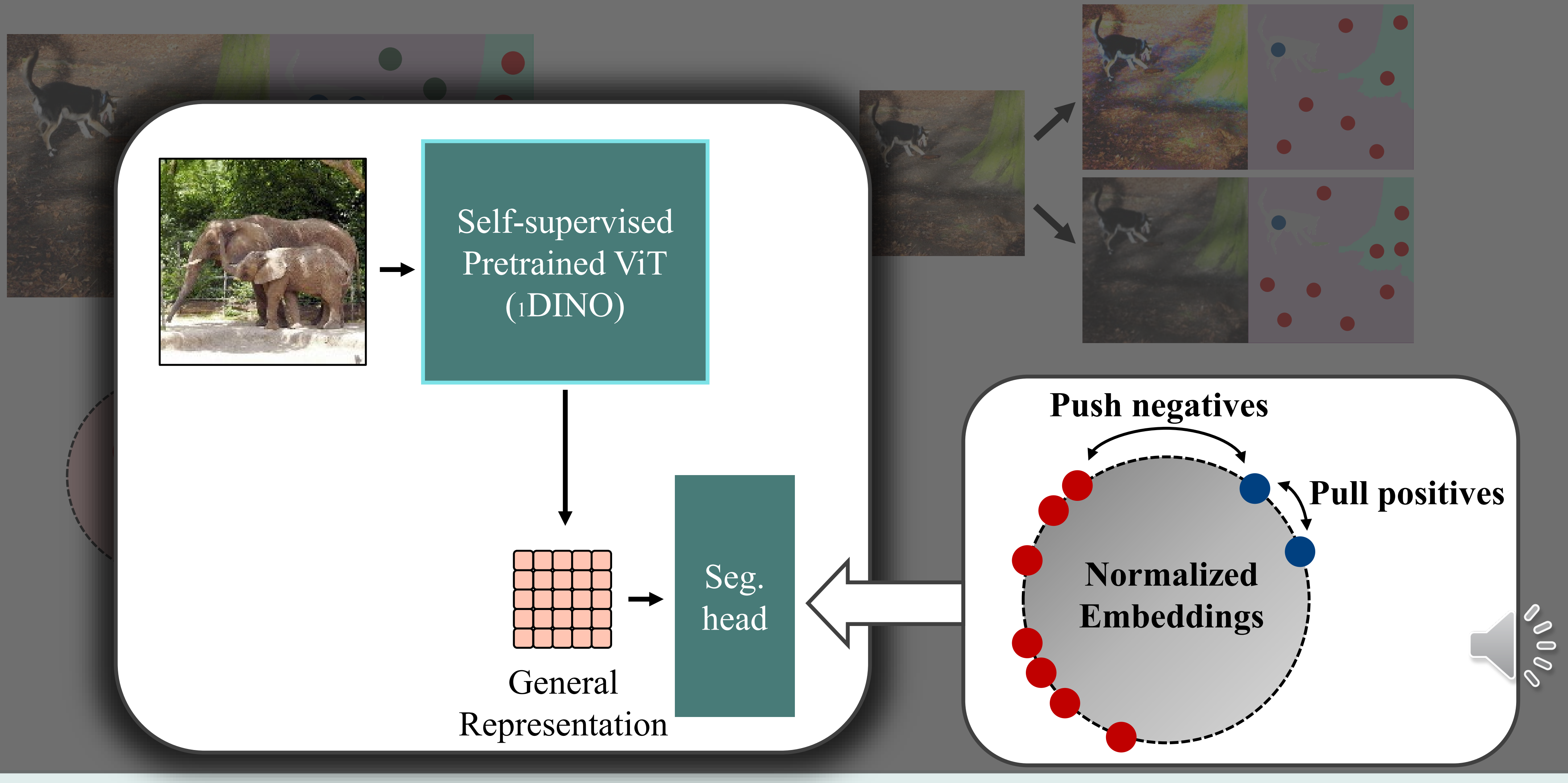




Common approaches

Clustering

Contrastive Learning



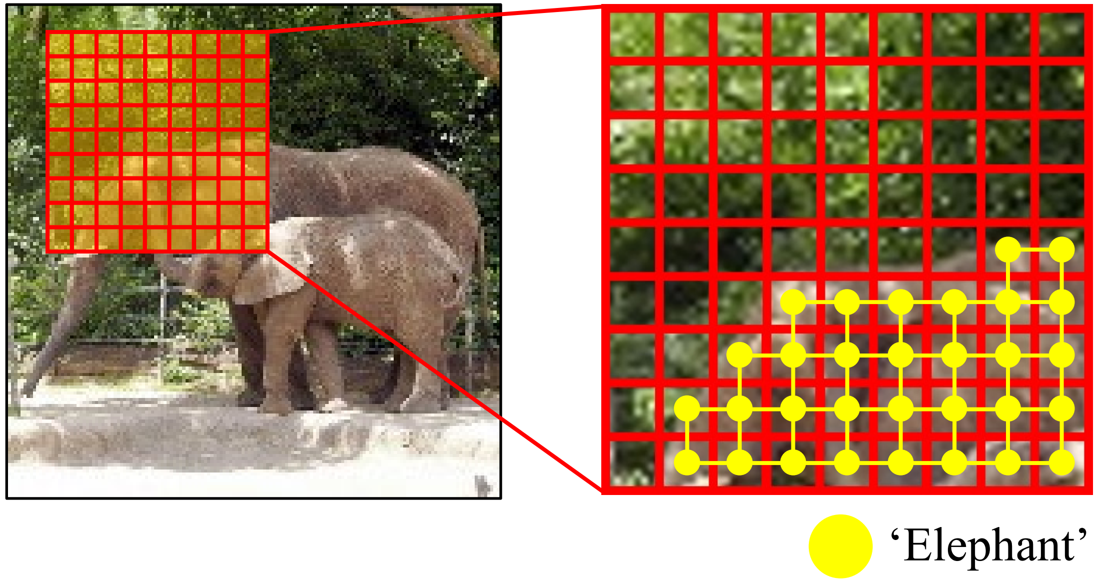
1: Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." ICCV 2021.

Motivation

Task-agnostic features from fixed pretrained model

- Task-agnostic features from self-supervised pretrained model (DINO) could be converted to segmentation features.
- However, relying solely on fixed pretrained model can be problematic since it is not specifically trained on segmentation task.

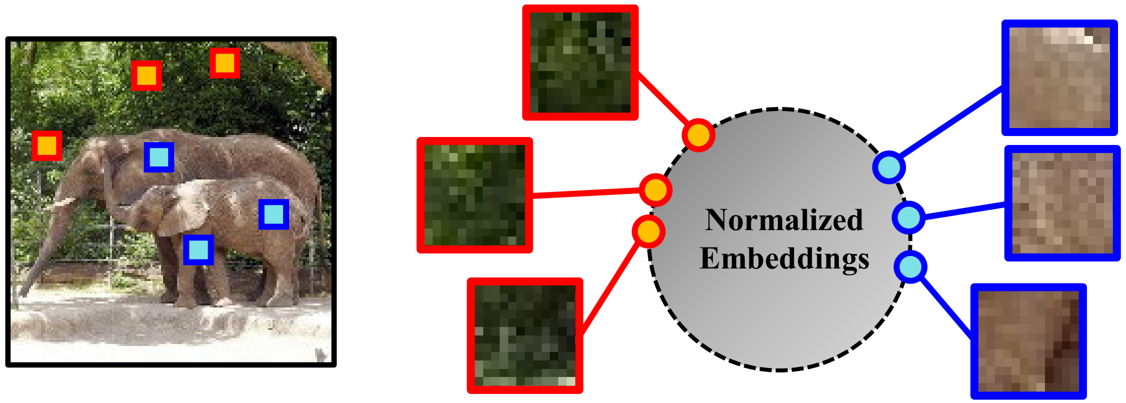
Local consistency



- Adjacent patches are highly likely to have analogous semantics, which could be a crucial clue for semantic segmentation.

Approach

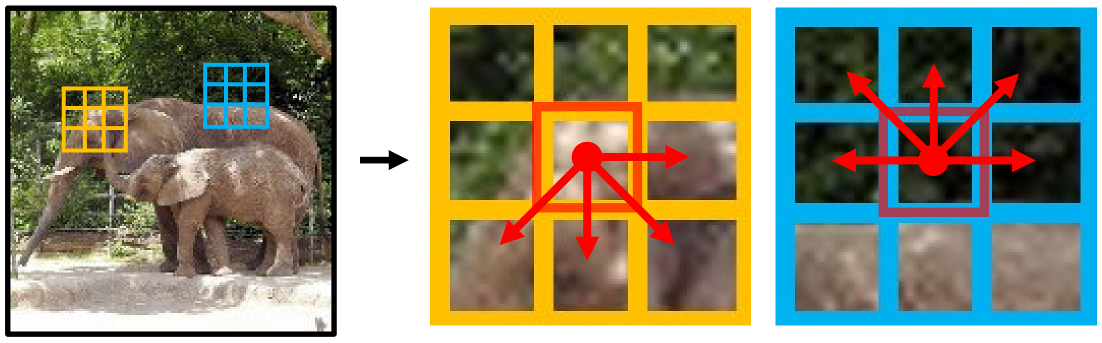
Global Hidden Positive (GHP)



Contrastive learning – pull positives & push negatives

- Due to absence of label, hidden positive patches (i.e., GHP) should be discovered.
- In addition to task-agnostic features, task-specific features should be used to discover GHP.

Local Hidden Positive (LHP)

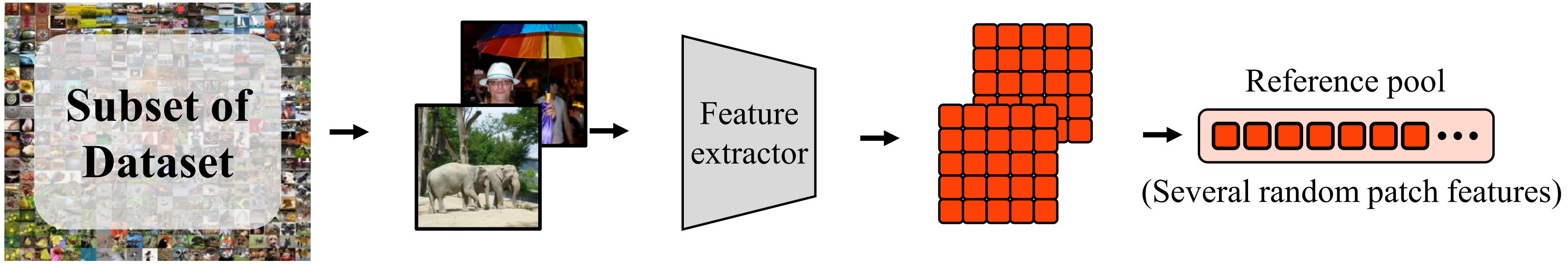


Propagate loss gradient

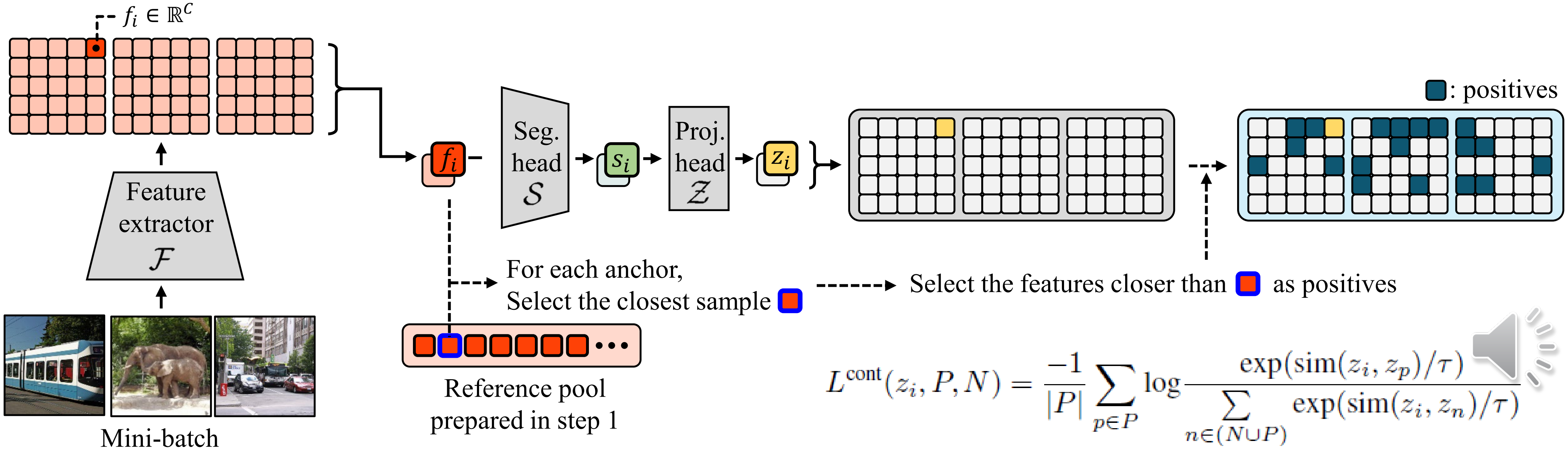
- Subset of surrounding patches which has high probability of having the same semantics with the anchor.
- Loss gradient is propagated to the LHP considering their equivalency.

Global Hidden Positive (GHP)

Stage 1. Build reference pool which stores prototypical patch features.



Stage 2. Discover GHP

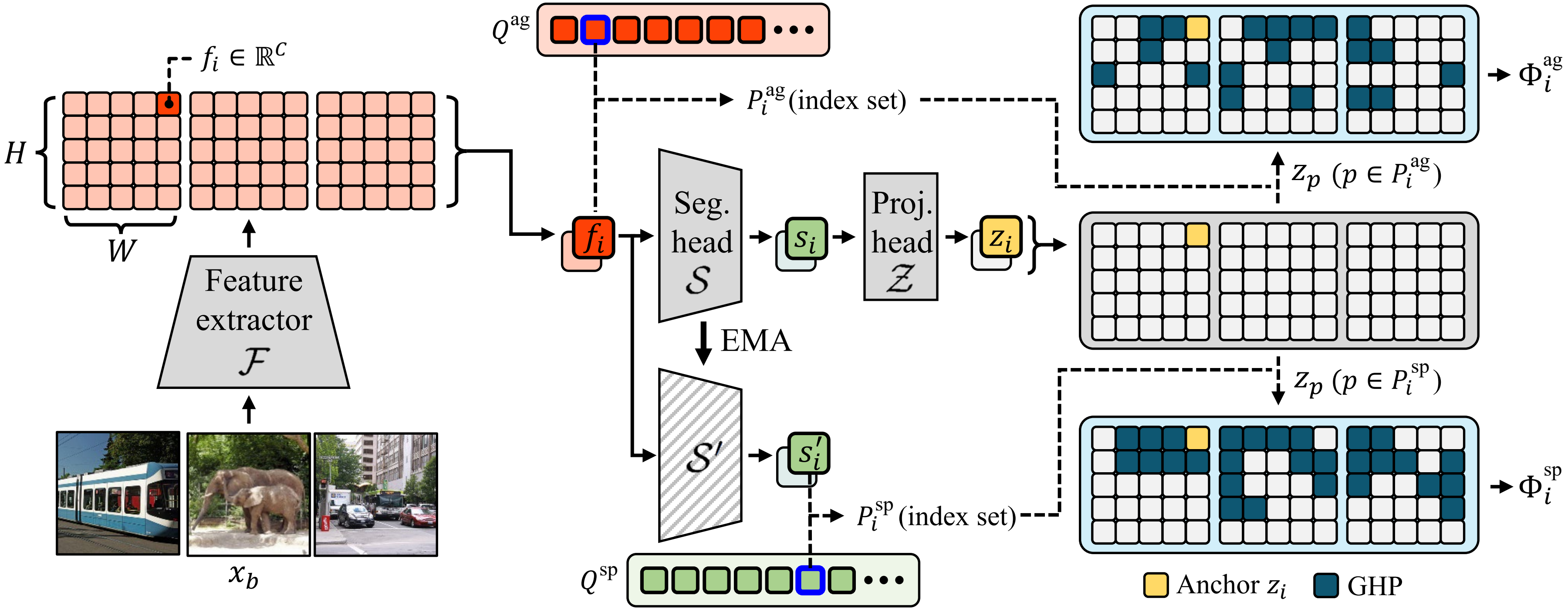


$$L^{\mathrm{cont}}(z_i, P, N) = \frac{-1}{|P|} \sum_{p \in P} \log \frac{\exp(\mathrm{sim}(z_i, z_p)/\tau)}{\sum_{n \in (N \cup P)} \exp(\mathrm{sim}(z_i, z_n)/\tau)}$$

Global Hidden Positive (GHP)

[Task-agnostic GHP] step 1-2 are carried out using patch features f_i.

[Task-specific GHP] step 1-2 are carried out using segmentation features s_i'.



$$L^{cont}(z_i, P, N) = \frac{-1}{|P|} \sum_{p \in P} \log \frac{\exp(\mathrm{sim}(z_i, z_p)/\tau)}{\sum_{n \in (N \cup P)} \exp(\mathrm{sim}(z_i, z_n)/\tau)}$$

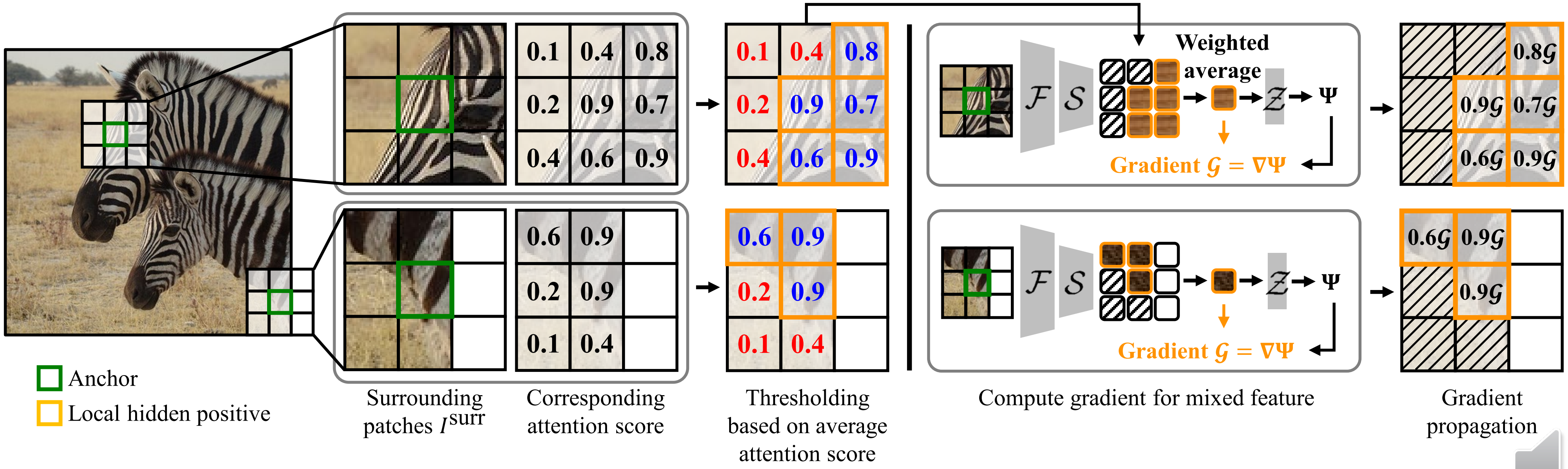
$$\Phi_i^{ag} = L^{cont}(z_i, P_i^{ag}, N_i^{ag})$$

$$\Phi_i^{sp} = L^{cont}(z_i, P_i^{sp}, N_i^{sp})$$

Training: $\Phi_i^{ag} + \lambda \Phi_i^{sp}$ (λ: $0 \longrightarrow 1$)

Local Hidden Positive (LHP)

- Propagate the loss gradient to adjacent patches.
  - Filter out the patches that have a lower attention score than the average attention score.
  - Propagate the gradient in proportion to the corresponding attention score.



Quantitative results

Experiment on COCO-stuff dataset

Method	Backbone	Unsupervised		Linear	
		Acc.	mIoU	Acc.	mIoU
DC [2]	R18+FPN	19.9	-	-	-
MDC [2]	R18+FPN	32.2	9.8	48.6	13.3
IIC [19]	R18+FPN	21.8	6.7	44.5	8.4
PiCIE [8]	R18+FPN	48.1	13.8	54.2	13.9
PiCIE+H [8]	R18+FPN	50.0	14.4	54.8	14.8
DINO	ViT-S/8	28.7	11.3	68.6	33.9
+ TransFGU [37]	ViT-S/8	52.7	17.5	-	-
+ STEGO [14]	ViT-S/8	48.3	24.5	74.4	38.3
+ Ours	ViT-S/8	**57.2**	**24.6**	**75.6**	**42.7**
DINO	ViT-S/16	22.0	8.0	50.3	18.1
+ STEGO [14]	ViT-S/16	52.5	23.7	70.6	34.5
+ Ours	ViT-S/16	**54.5**	**24.3**	**74.1**	**39.1**
SelfPatch	ViT-S/16	35.1	12.3	64.4	28.5
+ STEGO [14]	ViT-S/16	52.4	22.2	72.2	36.0
+ Ours	ViT-S/16	**56.1**	**23.2**	**74.9**	**41.3**

Experiment on Cityscapes dataset

Method	Backbone	Unsupervised		Linear	
		Acc.	mIoU	Acc.	mIoU
MDC [2]	R18+FPN	40.7	7.1	-	-
IIC [19]	R18+FPN	47.9	6.4	-	-
PiCIE [8]	R18+FPN	65.5	12.3	-	-
DINO	ViT-S/8	34.5	10.9	84.6	22.8
+ TransFGU [37]	ViT-S/8	77.9	16.8	-	-
+ Ours	ViT-S/8	**80.1**	**18.4**	**91.2**	**30.6**
DINO	ViT-B/8	43.6	11.8	84.2	23.0
+ STEGO [14]	ViT-B/8	73.2	**21.0**	90.3	26.8
+ Ours	ViT-B/8	**79.5**	18.4	**90.9**	**33.0**

Experiment on Potsdam-3 dataset

Method	Backbone	Unsup. Acc.
Random CNN [19]	VGG11	38.2
K-Means [27]	VGG11	45.7
SIFT [24]	VGG11	38.2
ContextPrediction [10]	VGG11	49.6
CC [18]	VGG11	63.9
DeepCluster [2]	VGG11	41.7
IIC [19]	VGG11	65.1
DINO	ViT-B/8	53.0
DINO + STEGO [14]	ViT-B/8	77.0
DINO + Ours	ViT-B/8	**82.4**

Qualitative results

- Experiment results on COCO-stuff dataset with DINO pretrained ViT-S backbone.

Further analysis

Ablation study

	GHP		LHP	SA	Unsupervised	
	TA	TS			Acc.	mIoU
(a)	✓	✓	✓	✓	57.2	24.6
(b)	✓	✓		✓	52.5	23.1
(c)	✓		✓	✓	55.0	19.1
(d)	✓			✓	49.3	20.1
(e)	✓	✓	✓		54.0	23.6
(f)					37.8	10.4

- TA and TS denote task-agnostic and task-specific, respectively.

⟷ Effect of Local Hidden Positive

⟷ Effect of task specific Global Hidden Positive

⟷ Comparison between ours and naïve implementation of unsupervised contrastive loss

Discovered GHP

- All anchors, reference points, and GHP sets have the same semantic labels.
- GHP selection process distinguishes the body parts in a more fine-grained manner.

Contributions

- We propose a novel method to discover semantically similar pairs, called global hidden positives, to explicitly learn the semantic relationship among patches for unsupervised semantic segmentation.

- We utilize the task-specific features from a model-in-training and validate the effectiveness of progressive increase of their contribution.

- A gradient propagation to nearby similar patches, local hidden positives, is developed to learn locality which is the most obvious clue in segmentation.

- Our approach outperforms existing state-of-the-art methods across extensive experiments.

Thank you

Leveraging Hidden Positives for Unsupervised Semantic Segmentation

Hyun Seok Seong, WonJun Moon, SuBeen Lee, and Jae-Pil Heo

Sungkyunkwan University