Self-Supervised Dense Representation Learning With Inter-Image Information

Julia Chae[2] Sandeep Routray[1,3] Amlan Kar[2,3] Sanja Fidler[2,3]
[1]IIT Kanpur [2]University of Toronto [3]Vector Institute
nayoung.chaeh@mail.utoronto.ca, sroutray@iitk.ac.in, {amlan, fidler}@cs.toronto.edu

Abstract

Dense self-supervised learning has shown promise for acquiring spatially rich visual representations without labels, yet most existing methods rely predominantly on intra-image signals and object priors, limiting their generalization to complex, scene-centric data. In this work, we propose an inter-image slot contrastive learning framework that enhances dense representation learning by comparing part-level features across different images. Our method builds on the dense clustering objective from [40] and introduces a novel slot-based inter-image contrastive loss, which mines semantically meaningful positive and negative image pairs using prototype-driven context vectors. We evaluate our approach on COCO and PASCAL unsupervised segmentation benchmarks, demonstrating improvements over intra-image-only baselines. Ablation studies further reveal the importance of inter-image positives and highlight that careful scheduling and weighting of the inter-image loss are essential for stable training. These results underscore the potential of inter-image supervision for learning more generalizable and part-aware dense representations.

1. Motivation

Self-supervised learning has significantly advanced representation learning by reducing dependence on labeled data. This allows models to acquire generic representations before fine-tuning for specific tasks with limited labeled data. However, most progress has been driven by image-level objectives, where models learn a single holistic representation per image. While effective for tasks such as classification and retrieval, this approach is insufficient for spatially diverse tasks like unsupervised image segmentation, where fine-grained, part-aware representations are crucial. Moreover, much of this success is built on well-curated, object-centric datasets such as ImageNet [6], which do not reflect the complexity of real-world scene-centric data. When instance discrimination is directly applied to such data, treating the entire scene as a single entity overlooks its rich internal structure—multiple objects, occlusions, and complex layouts—limiting the effectiveness of learned representations for dense prediction tasks.

A natural extension is dense representation learning, where models capture pixel-level features to support spatially aware tasks. While pixel-level learning methods [21, 29, 37] have shown strong performance in dense prediction, they primarily focus on local pixel relationships and fail to capture higher-level object relationships necessary for understanding complex scenes. Object-level representation learning has been explored as an alternative, but many existing approaches [10, 27, 32] rely on hand-crafted priors such as saliency estimation [26], object proposals [31, 34], or unsupervised clustering [14]. These priors introduce domain-specific biases that constrain representation learning and limit generalization across diverse real-world datasets.

To address these challenges, we propose a method that enhances self-supervised learning by incorporating inter-image relationships to improve dense representations. Instead of relying solely on augmentations of the same image, our approach compares learned features across different images using a dynamically updated queue to form meaningful positive and negative sample sets. Further, we integrate our method into the dense clustering pretext task in [40] to develop a finetuning procedure to learn richer, more transferable dense representations that effectively capture object relationships in complex scene-centric data. Our approach moves toward a more general and scalable self-supervised learning framework that does not rely on domain-specific priors, making it better suited for real-world dense prediction tasks. Our key contributions are as follows:

- We propose an inter-image slot contrastive learning framework for dense self-supervised representation learning, enabling part-level feature alignment across images.
- We develop a context-aware mining strategy that uses prototype-driven context vectors to select semantically meaningful positive and negative image pairs without requiring labels.
- We integrate our method into a dense clustering pretext task [40] and show that it can be applied on top of any pretrained model, making it flexible.

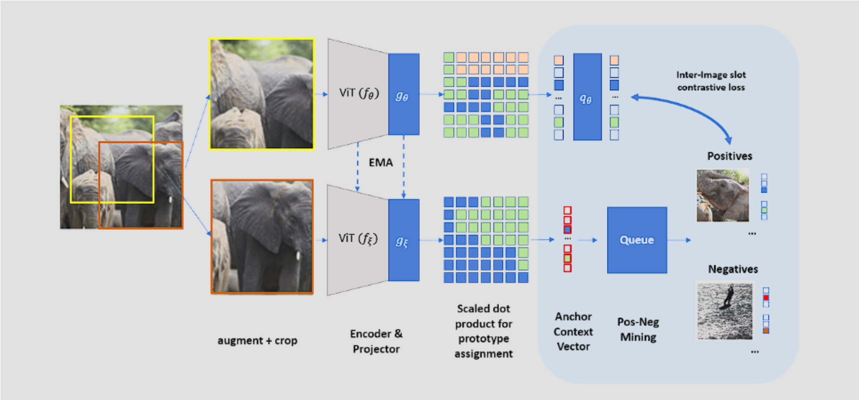

Figure 1. Illustration of our additive inter-image contrastive pipeline for investigating inter-image relationships in dense representation learning. To mine positive and negative samples, we compute a context vector from the anchor image's teacher features. Images with similar context vectors are selected as positives, while those with dissimilar context are treated as negatives.

- We conduct an extensive study on the role of mined inter-image positives and negatives, analyzing their individual contributions and interaction within the slot contrastive framework.
- We demonstrate that careful scheduling and weighting of the inter-image loss is crucial for stable training and improves performance on unsupervised segmentation benchmarks including COCO and PASCAL.

2. Related Works

2.1. Image-Level Representation Learning

Image-level representation learning treats the entire image as a single sample, learning holistic embeddings by comparing full-image representations. These methods have demonstrated strong performance on classification benchmarks such as ImageNet.

Intra-image supervision. Early approaches relied on reconstruction-based objectives, such as colorization [38], inpainting [20], image denoising [2, 28], or predicting pixel representations [4]. Other works introduced context-based reasoning using spatial jigsaw tasks [19], patch position prediction [7], or contrastive objectives across augmented views [8, 25]. Modern methods use self-distillation, where a student model is trained to match a momentum-updated teacher on different augmented views [3, 11].

Inter-image supervision. To enrich representations beyond intra-image cues, contrastive learning techniques have been developed that compare features across different images [5]. These methods define positive pairs using augmentations or nearest neighbors, while negatives are sampled either from the minibatch [5] or a memory queue [13]. Queue-based strategies such as MoCo [13] enable more diverse and stable learning, and extensions have explored various mining strategies [1, 9, 35, 39]. Our work builds on this line by applying inter-image supervision to dense representation learning rather than global image-level embeddings.

2.2. Dense Representation Learning

While image-level SSL excels at classification, it struggles with dense tasks such as detection or segmentation, which require spatially detailed and semantically structured features [12, 22]. This has led to increasing interest in dense representation learning.

Pixel-level contrastive learning. These methods extend instance discrimination to the pixel domain, matching pixel embeddings across views based on spatial alignment [24, 30, 36] or similarity in feature space [18]. While effective, they often rely on auxiliary image-level losses to stabilize training and prevent collapse.

Region-level learning. To capture more structured object-centric representations, several works introduce region-level objectives using heuristics such as saliency maps, contour detection, or region proposals [18]. Others discover semantic parts by maximizing mutual information between image patches [15] or individual pixels [14]. Recent works [17] have explored combining intra- and inter-image signals for dense learning, which closely relates to our approach. However, these often depend on handcrafted priors to define objectness.

Our work builds on SlotCon [33], which uses learned semantic slots instead of hand-designed priors to guide dense contrastive learning. We extend this by introducing an inter-image slot contrastive objective, encouraging feature consistency across instances of the same part in different images—leading to more robust and generalizable dense representations.

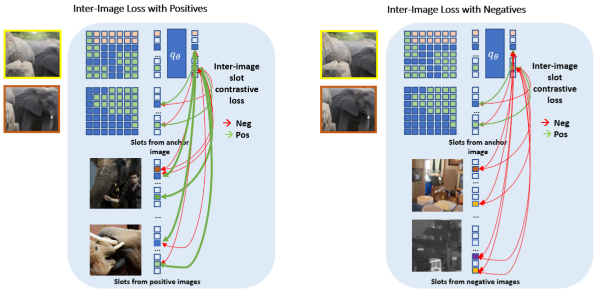
Inter-Image Loss with Positives
q_θ
Inter-image slot contrastive loss
Neg
Pos
Slots from anchor image
Slots from positive images
Inter-Image Loss with Negatives
q_θ
Inter-image slot contrastive loss
Neg
Pos
Slots from anchor image
Slots from negative images

Figure 2. Inter-image slot contrastive objectives based on mined positives and negatives. Positive images, selected for high semantic similarity to the anchor, contribute shared slots under diverse contexts to promote alignment across views. Negative images, chosen for low similarity, provide additional contrastive supervision with stronger control over semantic dissimilarity. Both losses are designed to complement the original intra-image slot contrastive objective.

3. Proposed Method

3.1. Baseline Framework Setup

To investigate the impact of inter-image relationships, we begin with a strong baseline: SlotCon [33], a state-of-the-art dense self-supervised learning framework that relies solely on intra-image contrastive supervision. In our experiments, we found that *initializing SlotCon with a DINOv2 backbone* (as in [40]) led to significantly better performance compared to training from scratch. We therefore adopt this configuration as our default baseline, illustrated in Figure 1.

The architecture consists of a student-teacher setup, where the teacher network ξ is updated using an exponential moving average of the student network θ. Both networks include an encoder (f_θ, f_ξ), a projector (g_θ, g_ξ), and a set of K learnable prototypes $S_\theta, S_\xi \in \mathbb{R}^{K \times D}$. Two augmented views v_1 and v_2 of an input image are passed through the encoder and projector to obtain spatial projections z_θ, $z_\xi \in \mathbb{R}^{HW \times D}$. These are l_2-normalized to yield $\bar{z}_\theta$ and $\bar{z}_\xi$, and similarly for the prototypes $\bar{S}_\theta$ and $\bar{S}_\xi$. The full training objective follows the formulation introduced in [33].

Object-level loss. This loss ensures that pixels belonging to the same object group are semantically consistent by matching ROI-aligned assignments between the teacher and student networks. First, we compute the teacher and student prototype assignments using their respective temperature parameters $\tau_t, \tau_s > 0$, a running logit center c, and an Invaug function to align augmented views:

$$\tilde{Q}_\xi = \text{Invaug}\left[\text{Softmax}_k\left((\bar{z}_\xi \cdot \bar{S}_\xi^\top - c)/\tau_t\right)\right] \tag{1}$$

$$\tilde{P}_\theta = \text{Invaug}\left[\text{Softmax}_k\left(\bar{z}_\theta \cdot \bar{S}_\theta^\top / \tau_s\right)\right] \tag{2}$$

We then apply cross-entropy loss to enforce assignment consistency between teacher and student across all spatial locations, symmetrically:

$$L_{\theta,\xi}^{\text{obj}} = \frac{1}{H \times W} \sum_{i,j} \Big[L^{\text{CE}}(\tilde{Q}_\xi^{(2)}[i,j], \tilde{P}_\theta^{(1)}[i,j]) + L^{\text{CE}}(\tilde{Q}_\xi^{(1)}[i,j], \tilde{P}_\theta^{(2)}[i,j]) \Big] \tag{3}$$

Intra-image contrastive loss. We perform a contrastive loss between pooled group-level feature vectors, referred to as *slots*. These are computed by first obtaining soft assignments:

$$A_\theta = \text{Softmax}_k\left(\bar{z}_\theta \cdot \bar{S}_\theta^\top / \tau_t\right) \tag{4}$$

$$W_\theta = \frac{1}{\sum_{i,j} A_\theta[i,j]} \sum_{i,j} A_\theta[i,j] \odot z_\theta[i,j] \tag{5}$$

To avoid comparing slots that are inactive (i.e., do not dominate any spatial location), we define a binary indicator:

$$1_\theta^k = \begin{cases} 1 & \text{if } \exists\,(i,j) \text{ such that } \arg\max_K\left(A_\theta[i,j]\right) = k \\ 0 & \text{otherwise} \end{cases}$$

The InfoNCE loss is applied across all valid slots, pulling together the same slot across views while pushing away others:

$$L_{\theta,\xi}^{\text{InfoNCE}}(W_\theta, W_\xi) = \frac{1}{K} \sum_{k=1}^{K} -\log \frac{1_\theta^k 1_\xi^k \exp\left(\bar{q}_\theta(w_\theta^k) \cdot \bar{w}_\xi^k / \tau_c\right)}{\sum_{k'} 1_\theta^k 1_\xi^{k'} \exp\left(\bar{q}_\theta(w_\theta^k) \cdot \bar{w}_\xi^{k'} / \tau_c\right)} \tag{6}$$

where $\bar{q}_\theta$ is the slot projector and τ_c is the temperature parameter.

Combined loss. The final SlotCon training objective combines the object-level loss and the intra-image slot contrastive loss:

$$L_{\text{baseline}} = L_{\theta,\xi}^{\text{obj}} + L_{\theta,\xi}^{\text{InfoNCE}} \tag{7}$$

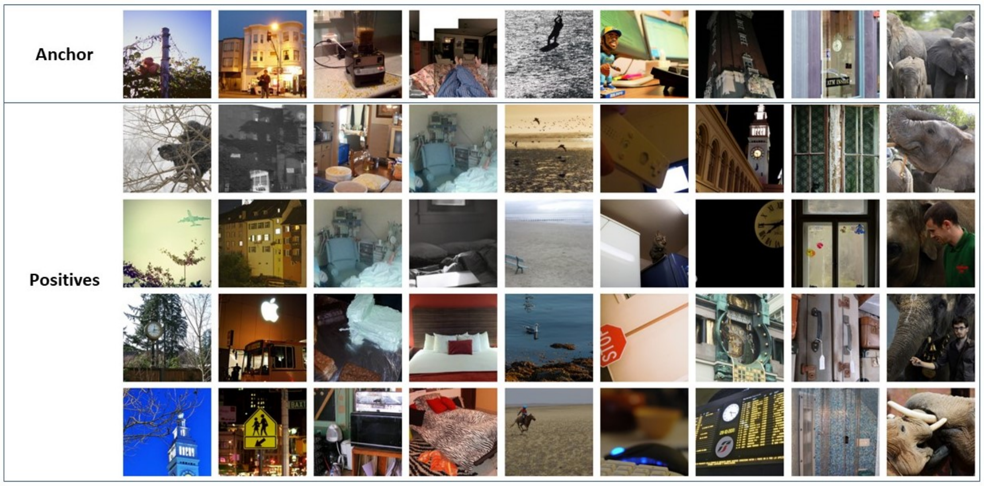

Figure 3. Example mined positives from contextual mining with pseudo-labels

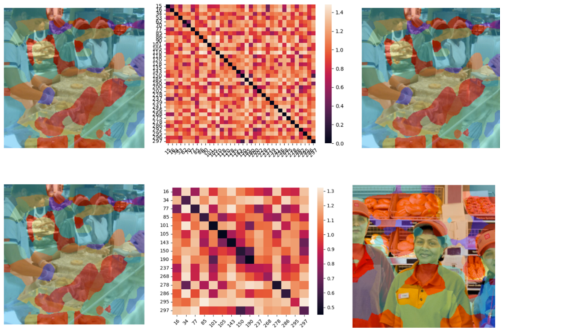

Figure 4. Similarity scores for shared slots across (Top): two views of the same image and (Bottom): two different images

3.2. Self-Supervised Inter-Image Mining

Why explicit inter-image supervision could help? The slot contrastive loss in Equation 6 can be extended to capture inter-image relationships by comparing slot features across different images, not just across views. To evaluate this, we visualized slot distances within and across images (Figure 4) and found that slots assigned to the same prototype across images are often closer than slots assigned to different prototypes within the same image. This indicates that intra-image training already induces inter-image semantic alignment—motivating the addition of explicit inter-image losses to further enhance representation quality.

A central challenge in inter-image contrastive learning is selecting semantically meaningful positive and negative images for each anchor, especially in the absence of labels. Motivated by prior work on contrastive mining [16, 23], we propose a self-supervised mining scheme that leverages the model's own prototype assignments as pseudo-labels.

Given the teacher feature map $\bar{z}_\xi \in \mathbb{R}^{HW\times d}$ and prototype matrix $\bar{S}_\xi \in \mathbb{R}^{K\times d}$, we compute the soft assignment of each pixel to the prototypes:

$$Z = \text{Softmax}\left(\frac{\bar{z}_\xi \cdot \bar{S}_\xi^\top}{\tau_m}\right), \quad Z \in \mathbb{R}^{HW\times K} \tag{8}$$

where τ_m is a temperature parameter controlling the sharpness of the assignments.

We summarize the prototype occurrence pattern for an image using a context vector $c \in \mathbb{R}^K$, defined as the sum of soft assignments over spatial positions:

$$c = \sum_{i=1}^{HW} Z_i \tag{9}$$

To measure the similarity between two images i and j, we use a Mahalanobis-style metric that considers co-occurrence patterns between prototypes. Let $\Sigma \in \mathbb{R}^{K\times K}$ be the covariance matrix and $\mu \in \mathbb{R}^K$ the mean vector computed over a queue of recent context vectors. Then, we define similarity as:

$$\text{Similarity}(c_i, c_j) = (c_i - \mu)^\top (\Sigma + \gamma I)^{-1} (c_j - \mu) \tag{10}$$

where γ is a small regularization constant.

This similarity score allows us to mine semantically relevant positive and negative images in an unsupervised manner. Images with high similarity to the anchor are selected as positives, while dissimilar ones are treated as negatives in the inter-image contrastive loss.

3.3. Inter-Image Loss Formulation for Negatives

We also extend the baseline intra-image contrastive loss by incorporating additional negatives from semantically dissimilar images mined from the queue. While the original

SlotCon formulation contrasts each anchor slot against non-matching slots from the same batch and different views, we augment this by including dominant prototypes from K_N mined negative images, identified using the context-based similarity metric.

Let $\mathcal{N}_{\text{intra}}$ denote the set of negatives from intra-image and batch views, and $\mathcal{N}_{\text{inter}}$ denote negatives from the mined images. The denominator in the InfoNCE loss becomes:

$$\mathcal{B} = \sum_{k' \in \mathcal{N}_{\text{intra}}} 1_\theta^k 1_\xi^{k'} \exp\left(\frac{\bar{q}_\theta(w_\theta^k) \cdot \bar{w}_\xi^{k'}}{\tau_c}\right) + \sum_{k'' \in \mathcal{N}_{\text{inter}}} 1_\theta^k 1_\xi^{k''} \exp\left(\frac{\bar{q}_\theta(w_\theta^k) \cdot \bar{w}_\xi^{k''}}{\tau_c}\right) \quad (11)$$

The updated contrastive loss with inter-image negatives is:

$$L_{\theta,\xi}^{\text{inter-neg}} = \frac{1}{K} \sum_{k=1}^{K} -\log \frac{1_\theta^k 1_\xi^k \exp\left(\bar{q}_\theta(w_\theta^k) \cdot \bar{w}_\xi^k / \tau_c\right)}{\mathcal{B}} \quad (12)$$

This formulation replaces the original intra-image contrastive loss, allowing us to isolate the impact of additional inter-image negatives.

3.3.1. Inter-Image Loss Formulation with Positives and Expanded Negatives

In our final formulation, each anchor image is paired with a mined positive image. The contrastive loss encourages similarity between corresponding slots across the pair while contrasting against an expanded set of negatives: intra-image, batch negatives, and additionally mined negatives from the queue.

Let P be the number of anchor-positive pairs, and K the number of slots per image. For each anchor-positive pair p, let $w_\theta^{k_p}$ be the anchor slot, $w_\xi'^{k_p}$ the matching positive slot, and $\mathcal{B}$ denote the total denominator including intra-image and inter-image negatives (as previously defined). The full inter-image contrastive loss with positives is:

$$L_{\theta,\xi}^{\text{inter}} = \frac{1}{K} \sum_{p=1}^{P} \sum_{k_p=1}^{K} -\log \frac{1_\theta^{k_p} 1_\xi^{k_p} \exp\left(\bar{q}_\theta(w_\theta^{k_p}) \cdot \bar{w}_\xi'^{k_p} / \tau_c\right)}{\mathcal{B}} \quad (13)$$

To ensure stable optimization, especially during early training when mining quality is poor, we apply a ramp-up weight $w_{\text{inter}} \in \mathbb{R}$. The total loss becomes:

$$L_{\text{total}} = L_{\theta,\xi}^{\text{obj}} + L_{\theta,\xi}^{\text{Inter-Neg}} + w_{\text{inter}} \cdot L_{\theta,\xi}^{\text{inter}} \quad (14)$$

This formulation unifies intra- and inter-image contrastive objectives, allowing the model to benefit from both local consistency and semantic alignment across diverse scenes.

4. Results

Training Details. We train our model on the COCO dataset using a ViT-Small architecture with patch size 16, initialized from DINO-pretrained weights. The training follows a student-teacher framework, where the teacher is updated via an exponential moving average of the student with a cosine schedule (starting at 0.9995 and annealed to 1.0). We train for 50 epochs using a batch size of 32 on 2 GPUs, with cosine learning rate schedules—starting at $1\text{e}{-4}$ for the projection head and $1\text{e}{-5}$ for the backbone. The projection head comprises three linear layers with GELU activations and outputs 256-dimensional features. All models are implemented in PyTorch and PyTorch Lightning, and clustering is performed using Faiss.

4.1. Evaluation Metrics

We assess dense representation quality via an unsupervised segmentation protocol based on overclustering, which directly operates on learned spatial features without additional supervision.

- **Feature Extraction.** Spatial features are extracted from the final layer of the backbone, discarding the projection head.
- **Overclustering.** K-Means clustering is applied to all spatial tokens from the validation set using $K = 500$ clusters to generate part-level pseudo-labels.
- **Label Matching.** Predicted clusters are matched to ground-truth classes by maximizing pixel-wise precision. The Hungarian algorithm ensures a permutation-invariant mapping.
- **Metric.** Mean Intersection over Union (mIoU) is computed between the aligned cluster map and ground-truth segmentation, averaged over five random K-Means seeds.

This protocol evaluates the semantic structure and spatial consistency of the learned representations without supervised fine-tuning.

4.2. Benchmarks

Table 1 presents a comparison of our method against a range of supervised and self-supervised baselines for transfer learning on unsupervised semantic segmentation. Results are reported in mIoU on PASCAL VOC 2012, COCO-Things, and COCO-Stuff.

Comparison with Supervised and Instance-Level Methods. Supervised models trained on large-scale datasets such as ImageNet (IN) and ImageNet21k (IN21) serve as upper bounds. Notably, our method—despite being trained in a self-supervised manner and only fine-tuned on COCO—achieves performance comparable to or better than supervised baselines on COCO-Things and COCO-Stuff. Compared to instance-level self-supervised methods like MoCo-v2 and SwAV, our approach provides substantial

Method	Train	PVOC12 (K=500)	COCO-Things (K=500)	COCO-Stuff (K=500)
Sup. ViT	IN + IN21	55.1	50.9	35.1
Sup. ResNet	IN	36.5	44.2	30.8
		instance-level		
MoCo-v2	IN	39.1	36.2	28.3
DINO	IN	17.4	23.5	32.1
SwAV	IN	35.7	37.3	33.1
		pixel/patch-level		
MaskContrast	IN + PVOC	45.4	37.0	25.6
DenseCL	IN	43.6	41.0	30.3
Ours (Best Neg.)	CC	47.5	**51.1**	**43.9**
Ours (Best Pos.)	CC	**48.7**	49.2	42.5

Table 1. Transfer learning results for semantic segmentation using KNN ($K = 500$). 'IN', 'IN21', 'CC', & 'PVOC' indicate training on ImageNet, ImageNet21k, COCO, & Pascal VOC training sets, respectively.

Experiment	No. of Mined Negatives	Queue size	PASCAL	COCO Stuff	COCO Things
Baseline	0		47.2	43.8	50.7
		Num Negatives Exp			
8 Rand negs	8	1024	47.5	44.1	50.3
15 Rand negs	15	1024	47.6	43.1	49.9
32 Rand negs	32	1024	46.9	43.45	49.8
64 Rand negs	64	1024	46.7	44.0	50.2
128 Rand negs	128	1024	47.5	43.9	51.1
		Queue Size Exp			
Queue size 512	15	512	46.8	43.9	49.6
Queue size 1024	15	1024	47.6	43.1	49.9
Queue size 2048	15	2048	47.5	44.0	50.2
Queue size 4096	15	4096	46.9	43.5	50.2

Table 2. Hyperparam. sweep on #mined negatives & queue size.

No. of Positive Images	w_{inter} Scheduling	Queue Size	PASCAL	COCO Stuff	COCO Things
0 (Baseline)	–	–	47.2	43.8	50.7
		Num Positive Experiments			
3	1	1024	46.1	40.1	46.4
6	1	1024	45.0	40.4	45.5
		w_{inter} Experiments (Varying w_{inter})			
6	1	1024	45.0	40.4	45.5
6	0.1	1024	**48.7**	42.3	49.2
6	0.01	1024	48.5	**42.5**	**49.7**

Table 3. Positive mining evaluation results with different weighting factors introduced from 10th epoch.

improvements, particularly on COCO-Things where object-level consistency is more critical.

DINO-V2 as Backbone. We fine-tune a strong pre-trained ViT backbone (DINO-V2) on COCO using our proposed inter-image contrastive framework. The standalone DINO baseline shows significantly lower performance compared to our method, especially on PASCAL VOC (17.4 vs. 48.7), demonstrating that the gains are not merely due to the backbone but rather the effectiveness of ourobject-level loss and inter-image loss formulation.

Effect of Inter-Image Loss. Both of our configurations—using best-performing inter-image negatives and positives—outperform pixel-level baselines such as DenseCL and MaskContrast, highlighting the importance of modeling relationships across images. Our best negative mining configuration achieves the highest performance on COCO-Things (51.1 mIoU), while the best positive mining configuration achieves the strongest result on PASCAL VOC (48.7 mIoU). These results indicate that inter-image relationships improve generalization across datasets with diverse scenes and object distributions.

4.3. Effect of Mined Negatives and Queue Size

Table 2 summarizes the impact of varying both the number and hardness of negatives in the inter-image contrastive loss. Increasing the number of negatives or selecting harder negatives (top 2% most similar images from the memory queue) resulted in negligible performance gains. These findings suggest that SlotCon's original slot-contrastive loss—based on negatives from other images within the batch—already provides a strong supervisory signal, and that supplementing it with additional or harder negatives offers limited improvement.

4.4. Effect of Positive Sampling & Loss Weight Scheduling

Table 3 demonstrates that incorporating additional positive samples and enforcing prototype consistency with the anchor improves representation learning. To ensure the reliability of positive samples, we adopted a stringent similarity threshold (<1% to the anchor, instead of the conventional 25%). The consistency loss was introduced only in later training stages, once prototypes had stabilized, as early inclusion was found to degrade performance due to noisy feature representations. Additionally, we applied a small weighting factor to this loss to prevent interference with the primary objective and to reduce noise.

5. Conclusion

In summary, our proposed method leveraging inter-image information shows promise for learning more generalizable & part-aware dense representations. Our findings underscore the utility of inter-image supervision & the practicality of fine-tuning existing pretrained models. We demonstrate that object-centric contrastive learning can be effectively adapted for scene-centric datasets without requiring full retraining, making it broadly applicable to downstream dense prediction tasks.

References

[1] Mehdi Azabou, Mohammad Gheshlaghi Azar, Ran Liu, Chi-Heng Lin, Erik C Johnson, Kiran Bhaskaran-Nair, Max Dabagia, Bernardo Avila-Pires, Lindsey Kitchell, Keith B Hengen, et al. Mine your own view: Self-supervised learning through across-sample prediction. *arXiv preprint arXiv:2102.10106*, 2021. 2

[2] Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In *International Conference on Machine Learning*, pages 517–526. PMLR, 2017. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[4] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[7] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 2

[8] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014. 2

[9] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9588–9597, 2021. 2

[10] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals, 2021. 1

[11] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[12] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019. 2

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[14] Olivier J Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. In *European Conference on Computer Vision*, pages 123–143. Springer, 2022. 1, 2

[15] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019. 2

[16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 4

[17] Xiaoni Li, Yu Zhou, Yifei Zhang, Aoting Zhang, Wei Wang, Ning Jiang, Haiying Wu, and Weiping Wang. Dense semantic contrast for self-supervised visual representation learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1368–1376, 2021. 2

[18] Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *arXiv preprint arXiv:2011.13677*, 2020. 2

[19] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pages 69–84. Springer, 2016. 2

[20] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[21] Pedro O. Pinheiro, Amjad Almahairi, Ryan Y. Benmalek, Florian Golemo, and Aaron Courville. Unsupervised learning of dense visual representations, 2020. 1

[22] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *Advances in Neural Information Processing Systems*, 33:3407–3418, 2020. 2

[23] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020. 4

[24] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatially consistent representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1153, 2021. 2

[25] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020. 2

[26] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10052–10062, 2021. 1

[27] José-Fabian Villa-Vásquez and Marco Pedersoli. Unsupervised object discovery: A comprehensive survey and unified taxonomy, 2024. 1

[28] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 2

[29] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[30] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 2

[31] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems*, 34:22682–22694, 2021. 1

[32] Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping, 2022. 1

[33] Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. *Advances in neural information processing systems*, 35:16423–16438, 2022. 2, 3

[34] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. *Advances in Neural Information Processing Systems*, 34:28864–28876, 2021. 1

[35] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Delving into inter-image invariance for unsupervised visual representations. *International Journal of Computer Vision*, 130(12):2994–3013, 2022. 2

[36] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021. 2

[37] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning, 2021. 1

[38] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016. 2

[39] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6002–6012, 2019. 2

[40] Adrian Ziegler and Yuki M. Asano. Self-supervised learning of object parts for semantic segmentation, 2022. 1, 3