Pixel-weighted Domain Adaptation for Agricultural Segmentation

Shunta Kimura*, Handie Shao*, Shogo Matsumoto†,
Daiki Yamada†, Toshihiro Kitajima†, and Hideki Nakayama*
* The University of Tokyo, Japan
E-mail: 4018940347@g.ecc.u-tokyo.ac.jp, {shao, nakayama}@nlab.ci.i.u-tokyo.ac.jp
† Kubota Corporation, Japan
E-mail: {shogo.matsumoto, daiki.yamada3, toshihiro.kitajima }@kubota.com

***Abstract*—Semantic segmentation plays a critical role in autonomous driving, but collecting labeled data in real-world agricultural environments is labor-intensive and costly. While transfer learning from existing datasets is a promising solution, most benchmark datasets such as Cityscapes are designed for urban scenes, making them poorly suited for rural domains due to domain gaps and label definition mismatches. To address these challenges, we investigate how supervised domain adaptation (DA) techniques can enhance model transfer beyond naive fine-tuning on limited target data. Our framework incorporates a pixel-level domain discriminator and both pixel-wise and class-wise importance weighting to explicitly reduce domain and label distribution discrepancies. We also explore an entropy-based label alignment module to mitigate label mismatch, though it was ultimately excluded from the final configuration due to limited effectiveness. Experiments show that our method improves mIoU by +13.8% over the target-only baseline, and by +2.4% over naive fine-tuning using both source and target data. These results demonstrate that properly designed supervised DA strategies can offer significant benefits over simple fine-tuning, especially under severe domain shifts.**

I. Introduction

Semantic segmentation is essential for autonomous driving, providing pixel-level scene understanding for navigation and safety. While most research focuses on structured urban scenes, rural and agricultural environments remain underexplored despite growing importance. This is particularly important in many developed countries, where aging populations and labor shortages are accelerating demand for agricultural automation.

In these agricultural environments, unique challenges emerge, such as unpaved roads, dense vegetation, domain-specific objects, and diverse weed types. Yet, they also share visual elements with urban scenes, including roads, sky, and buildings. This partial overlap makes transfer learning from urban datasets promising for agricultural segmentation. Datasets like Cityscapes [1], BDD100K [2], and Mapillary Vistas [3] have advanced urban segmentation and can provide useful prior knowledge. However, three key issues hinder direct transfer: domain shift, label mismatch, and class imbalance.

Domain shift arises from differences in lighting, texture, object distribution, and layout. As shown in Fig. 1, farm scenes feature overgrown vegetation, unpaved roads, and wide-open fields, which are not typically present in urban environments.

In addition, semantic labels differ between domains. Some urban classes such as "sidewalk" are irrelevant in rural settings, while new classes like *weeds* or *field* appear only in the target domain, leading to label mismatch.

Moreover, class imbalance is pronounced, as rare but important classes such as *person* or *vehicle* are significantly underrepresented in the target domain, requiring imbalance-aware training strategies.

To address these challenges, we propose a unified framework that enhances supervised domain adaptation by tackling domain shift and class imbalance. Our method combines pixel-level adversarial training and both pixel- and class-wise importance weighting. We also explored an entropy-based label alignment module to handle label mismatch, but excluded it from the final model due to limited effectiveness.

Experiments show that our method improves mIoU by +13.8% over the target-only baseline and by +2.4% over naive fine-tuning, demonstrating the value of supervised DA under limited target data.

Our main contributions are:

- A unified DA framework that outperforms naive fine-tuning by addressing domain shift and class imbalance.
- Integration of pixel-level adversarial training and multi-scale importance weighting.
- Exploration of entropy-based label alignment, excluded from the final model due to limited gains.
- Strong performance using only 481 labeled target images, enabling practical deployment in data-scarce domains.

Our framework offers an efficient solution for semantic segmentation in unstructured or underrepresented domains.

II. Related Work

A. Domain Adaptation

Domain adaptation for semantic segmentation typically aligns source and target domains in the input, feature, or output space. Feature-level adversarial methods such as DANN [4] aim to make feature distributions domain-invariant, but global-level alignment often causes negative transfer, especially for rare classes. To address this, finer-grained approaches such as pixel-level discriminators have been introduced, as in PixDA

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

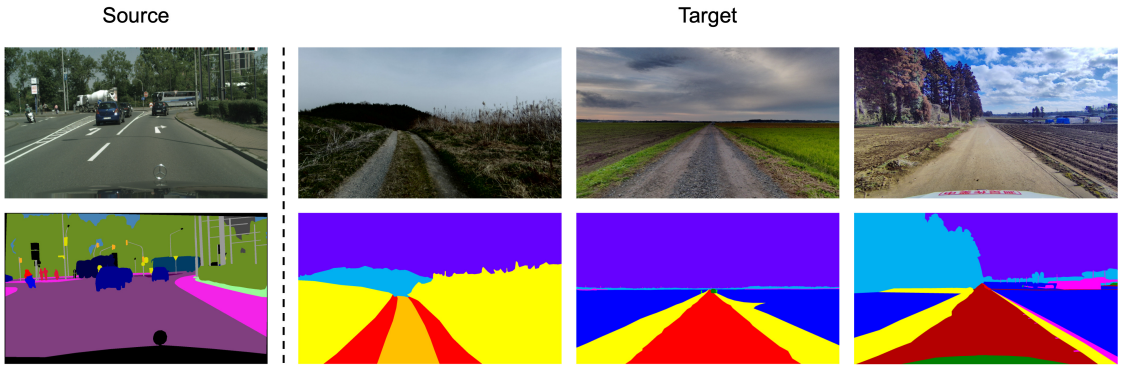

Fig. 1. Example images from the source and target domains. The leftmost column shows a sample from the Cityscapes dataset (urban streets), while the remaining columns present samples from our farm road dataset (rural environments). There exists a significant domain gap between the two: not only in visual appearance and scene layout, but also in the set of semantic classes. While some classes such as *road* or *vehicle* are shared, others are unique to the target domain, such as *weeds* or *field*. This highlights the challenges of cross-domain adaptation for agricultural applications.

[5], which reweights adversarial loss based on pixel uncertainty and rarity.

Our method follows this line by employing a pixel-wise discriminator with importance weighting, allowing more stable adaptation under class imbalance.

B. Label Alignment

Most domain adaptation works assume shared label sets. However, in open-set or partially overlapping scenarios, this assumption does not hold. Some studies [6], [7] use clustering or label embeddings to bridge label gaps. Notably, entropy-based methods [7] minimize uncertainty in similarity distributions between image features and label embeddings.

We also explore an entropy-based label alignment module that aims to handle class mismatch across domains, though it was ultimately excluded from the final model due to limited effectiveness.

C. Label Imbalance

Semantic segmentation is sensitive to label imbalance, where dominant classes (e.g., road, sky) overshadow rare ones (e.g., person, vehicle). While sampling-based solutions exist, recent works [8] emphasize pixel-wise weighting based on class rarity and confidence.

However, most existing approaches apply such reweighting only to the segmentation loss. In contrast, we extend the use of importance weights to both adversarial and label alignment losses. This unified weighting strategy improves the segmentation performance of rare classes while mitigating overfitting, especially under limited target data scenarios.

III. Method

A. Problem Definition

We consider the task of semantic segmentation under the supervised domain adaptation setting, where the goal is to train a model that performs well on a target domain using labeled data from both source and target domains.

Let X be the space of RGB images composed of pixels $i \in I$, and Y the space of corresponding semantic masks with class labels from a set C. We are given two labeled datasets: source domain samples $X_s = \{(x^s, y^s)\}$ and target domain samples $X_t = \{(x^t, y^t)\}$, where $x^s, x^t \in X$ and $y^s, y^t \in Y$. The class sets in the source and target domains are denoted as C_s and C_t, which may partially overlap or differ, i.e., $C_s \neq C_t$.

Our goal is to learn a segmentation model $f_\theta : X \to \mathbb{R}^{|I| \times |C_t|}$ that predicts pixel-wise class probabilities on target images. The model consists of a shared encoder f_θ^{enc} parameterized by θ^{enc}, and domain-specific segmentation heads $f_\theta^{\text{seg},s}$ and $f_\theta^{\text{seg},t}$ for source and target domains respectively. The overall prediction is obtained as:

$$f_\theta(x) = \begin{cases} f_\theta^{\text{seg},s}(f_\theta^{\text{enc}}(x)), & \text{if } x \in X_s, \\ f_\theta^{\text{seg},t}(f_\theta^{\text{enc}}(x)), & \text{if } x \in X_t. \end{cases} \tag{1}$$

We use the encoder features not only for segmentation but also as input to auxiliary components such as a pixel-wise domain discriminator and an entropy-based label alignment module. For a pixel i and class c, the model's predicted probability is denoted as $p_i^c(x) = f_\theta(x)[i, c]$.

B. Overview of the Framework

To address the challenges of domain shift and label mismatch described above, we propose a unified framework composed of the following components:

- **Segmentation Module:** Includes a shared encoder f^{enc} and domain-specific segmentation heads f^s and f^t. Class-weighted losses are applied to mitigate label imbalance in each domain.
- **Pixel-Wise Domain Discriminator:** A discriminator D operating at the pixel level encourages the encoder to produce domain-invariant features. Importance weighting is incorporated to reduce negative transfer and to promote rare-class adaptation.
- **Entropy-Based Label Alignment (Optional):** When source and target class sets differ, an entropy-based module helps align semantics across domains by minimizing prediction uncertainty. Pixel-wise importance weights further emphasize underrepresented regions.

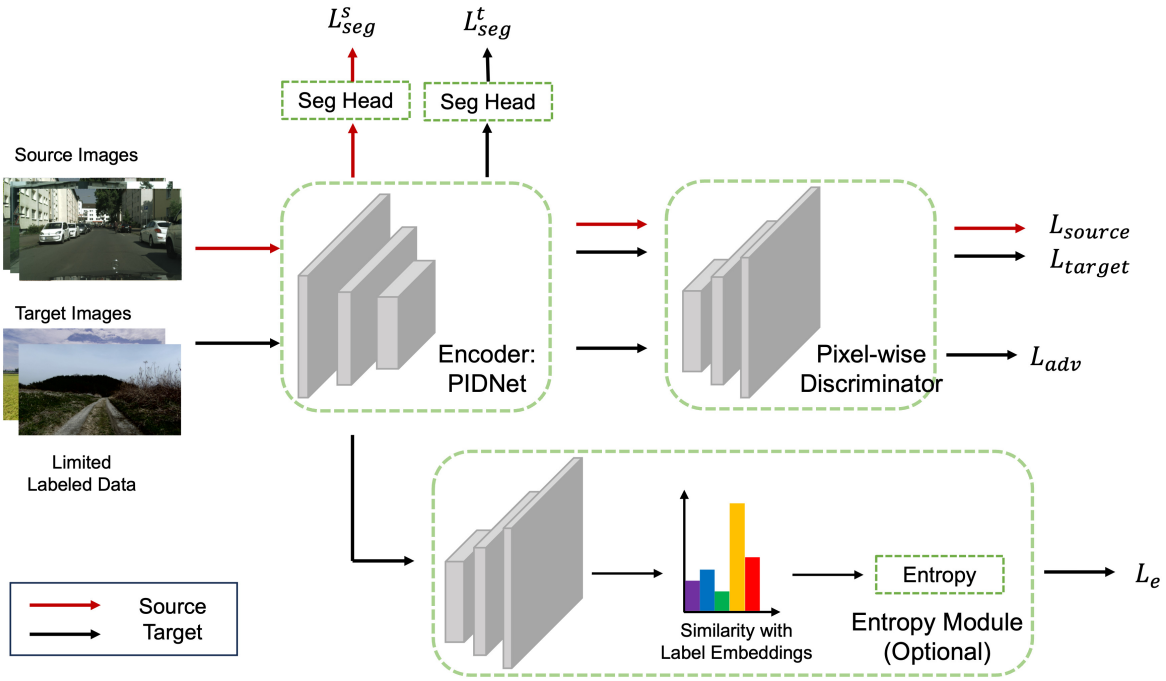

Fig. 2. Main structure of our proposed model, consisting of three modules for semantic segmentation, domain adaptation, and optional label alignment.

We describe each of these components in detail in the following sections.

C. Basic Network: PIDNet [9]

We adopt PIDNet [9] as the shared encoder for its efficiency and strong real-time segmentation performance. It processes both source and target images and outputs feature maps used by the segmentation heads, domain discriminator, and entropy module. Each domain has its own segmentation head, supervised by corresponding ground-truth labels.

To train the segmentation heads, we use the Online Hard Example Mining (OHEM) Cross-Entropy loss [10], which focuses on low-confidence pixels to emphasize challenging regions such as boundaries and rare classes. To further address class imbalance, we apply class-wise weights based on the inverse log-frequency of target classes, smoothing the impact of extremely rare or dominant categories. These strategies enhance the model's ability to learn from underrepresented regions, which is crucial in cross-domain settings. We denote the segmentation losses for the source and target domains as $\mathcal{L}^s_{\text{seg}}$ and $\mathcal{L}^t_{\text{seg}}$, respectively.

D. Pixel-Wise Discriminator for Domain Adaptation

Following prior work on fine-grained domain adaptation [5], we incorporate a pixel-wise adversarial discriminator D , which estimates the probability that a given pixel originates from the source domain.

The adversarial loss is defined as:

$$\mathcal{L}_{\text{adv}} = -\frac{1}{|I|}\sum_{i\in I} S_i B_i \log D_i, \tag{2}$$

where $D_i = D(x)_i \in (0,1)$ denotes the predicted probability that pixel i belongs to the source domain.

The importance weights S_i and B_i are given by:

$$S_i = -y_i \log p_i(y_i), \quad B_i = 1 - \frac{1}{|I|}\sum_{j\in I} \mathbb{1}[y_j = y_i]. \tag{3}$$

Here, S_i represents the uncertainty (negative log-probability of the predicted class), and B_i reflects the inverse frequency of the ground-truth class.

The composite weight $S_i B_i$ increases the contribution of uncertain or rare-class pixels. This reduces overfitting to majority classes and enhances domain alignment for underrepresented categories.

E. (Optional) Entropy Module for Label Alignment

Although this module was not included in the final model due to limited effectiveness, we explored an entropy-based approach to address label mismatch by leveraging external semantic knowledge. Specifically, we use CLIP's [11] text encoder to extract fixed semantic embeddings for each class label in the source and target domains. These embeddings are then compared with pixel features to guide cross-domain alignment.

Let $c^k_s = \text{CLIP}_{\text{text}}(y^k_s)$ and $c^k_t = \text{CLIP}_{\text{text}}(y^k_t)$ denote the text embeddings of the k-th source and target class labels, respectively. Following the formulation of [7], the encoder outputs are projected into the same embedding space, and cosine similarities are computed as:

$$[v_{ts}]_k = \phi(\mathcal{E}(f^{enc}_\theta(x_t)), c^k_s), \quad [v_{tt}]_k = \phi(\mathcal{E}(f^{enc}_\theta(x_t)), c^k_t). \tag{4}$$

Method	Target	Source	Weighting	Discriminator	Entropy	mIoU ↑
Target-only	✓					26.43
Multi-task(fine-tuning)	✓	✓				37.91
+Weighting	✓	✓	✓			38.38
+Discriminator	✓	✓	✓	✓		**40.26**
Entropy-only	✓	✓			✓	31.12
Full Model	✓	✓	✓	✓	✓	32.50

TABLE I
ABLATION STUDY OF OUR METHOD. EACH COMPONENT IS INCREMENTALLY ADDED TO THE BASE FINE-TUNING MODEL.

Class	Target-only	Best Model
Other Obstacle	23.9	**25.5**
Road	**15.1**	13.5
Dirt road	0.3	**6.3**
Road weed	8.7	**9.2**
Other weed	37.7	**49.8**
Person	22.3	**45.6**
Vehicle	2.6	**31.2**
Trees	58.9	**74.7**
Field	22.5	**55.8**
Sky	72.4	**90.9**
Mean IoU	26.4	**40.3**

TABLE II
PER-CLASS IOU (%) COMPARISON BETWEEN TARGET-ONLY AND OUR PROPOSED METHOD ON THE AGRICULTURAL DATASET.

Here, $\mathcal{E}(\cdot)$ denotes a lightweight projection network that maps encoder outputs to the label embedding space. The similarity function $\phi(\cdot,\cdot)$ is implemented as a 1×1 convolution initialized with L2-normalized CLIP embeddings and kept fixed during training.

We then compute the entropy of the similarity distributions to encourage confident alignment, where $\sigma(\cdot)$ denotes the softmax operator:

$$E_i = H(\sigma([v_{ts}])) + H(\sigma([v_{tt}])). \tag{5}$$

The final entropy loss is defined as:

$$L_{\text{e}} = -\frac{1}{|I|}\sum_{i\in I} S_i B_i \log E_i. \tag{6}$$

This encourages the model to align pixel-level features with semantically relevant labels from both domains, while the weighting terms S_i and B_i reduce the impact of noisy or ambiguous regions.

F. Total Loss Function

The final training objective integrates segmentation, adversarial, and (optionally) entropy-based alignment losses:

$$\mathcal{L} = \mathcal{L}^s_{\text{seg}} + \mathcal{L}^t_{\text{seg}} + \alpha\mathcal{L}_{\text{adv}} + \beta\mathcal{L}_{\text{ent}}.$$

This formulation allows for flexible balancing between segmentation, domain adaptation, and label alignment.

IV. EXPERIMENT

A. Datasets

We use two datasets for domain adaptation: Cityscapes [1] as the source domain and a custom farm road dataset from Japan as the target domain.

Cityscapes provides 2,975 finely annotated urban street images with 19 semantic classes. We use the entire training split. The **target dataset** consists of 880 manually labeled images from rural farm roads, covering 10 classes including *road*, *weeds*, *field*, *vehicle*, and *person*. To evaluate generalization, we split the dataset by location: 481 images from 2 areas are used for training, and 399 images from 4 unseen areas are used for evaluation.

All images are resized to 1024×512. The target domain shows strong class imbalance, especially for rare classes like *person* and *vehicle*, which motivates our imbalance-aware adaptation strategy.

B. Experiment Setting

1) Model Selection: Among the available PIDNet variants, we select PIDNet-M, which offers a good balance between segmentation accuracy and model size (34.4M parameters).

2) Training: We use the publicly available PIDNet-M model pre-trained on the Cityscapes dataset to provide strong performance on the source domain. Then, the full domain adaptation framework is trained using Cityscapes as the source domain and the farm road dataset as the target domain for 15,000 iterations. Stochastic gradient descent (SGD) is applied to the encoder, while Adam is used for optimizing the discriminator and entropy modules. Learning rates are set to 2.5×10^{-4} (SGD) and 10^{-5} (Adam), with a weight decay of 5×10^{-4}. Momentum is set to 0.9 for SGD and $\{0.9, 0.99\}$ for Adam. Loss weights α and β are set to 0.1. Training is conducted on a single NVIDIA RTX A6000 GPU with a batch size of 4.

C. Results

1) Ablation Study and Results: Table I shows an ablation study evaluating the contribution of each module in our framework. Starting from a simple fine-tuning baseline (multi-task), we observe a significant improvement from 26.43% to 37.91% mIoU by leveraging both source and target labels. Incorporating pixel-wise weighting based on source-target alignment further boosts performance to 38.38%, highlighting the importance of addressing class imbalance. Adding a domain

discriminator improves the performance to 40.26%, confirming the benefit of adversarial alignment.

It is notable that both the entropy-only setting (31.12%) and the full model (32.50%) underperform compared to the best-performing combination. We attribute this limitation to the nature of CLIP-based label embeddings, which are derived from large-scale, generic image-text pairs and are not specialized for the domain-specific visual characteristics of our dataset. For example, *road* and *dirt road* are semantically similar (cosine similarity ≈ 0.9), but their appearances differ significantly in rural environments. Since such semantically similar labels tend to be closely distributed in the embedding space regardless of visual diversity, this may lead to feature space confusion or clustering, especially under domain shift.

2) Per-class IoU Results: Table II presents per-class IoU improvements. Significant gains are observed in rare or domain-shifted categories, such as *vehicle* ($2.6 \rightarrow 31.2$), *person* ($22.3 \rightarrow 45.6$), and *field* ($22.5 \rightarrow 55.8$), confirming the effectiveness of our approach in addressing class imbalance and label mismatch.

We hypothesize that the substantial improvement in *vehicle* and *person* is due to effective transfer from the source domain, which contains a large number of such instances, despite their sparsity in the target domain. In contrast, classes like *other weed* and *field*, which lack direct counterparts in the source dataset, benefit from better feature separation enabled by the proposed modules.

On the other hand, lower improvements in *dirt road* and *road weed* may stem from the absence of corresponding source labels and the semantic proximity to other visually similar classes (e.g., *road* and *weed*), leading to potential confusion. Notably, although *dirt road* and *road weed* improve slightly, *road* IoU decreases ($15.1 \rightarrow 13.5$). This may be due to visual domain bias: Cityscapes' *road* depicts clean urban pavements, whereas target paved roads are often narrow and soiled or partially covered by mud or vegetation, which can cause ambiguous segments to be misclassified as dirt-like classes.

V. Limitations and Future Work

While our method demonstrates significant improvements, certain limitations remain. First, transferring knowledge from the source domain is still challenging for classes with large visual gaps or without direct label correspondences in the source dataset. In particular, performance gains are limited for semantically similar classes such as *road* and *dirt road*, likely due to label confusion caused by overlapping embeddings.

For future work, improving label alignment remains a key direction. This includes designing label embeddings that are not only semantically meaningful but also visually well-distributed within each domain. Integrating visual prototypes or embedding adaptation techniques could further reduce cross-domain confusion. In addition, combining our framework with test-time adaptation or self-supervised learning may enhance robustness in dynamic and unseen environments.

VI. Conclusion

We addressed the challenges of domain adaptation for semantic segmentation in farm road environments, where training data is scarce and issues such as class imbalance and label mismatch are prominent. Our method incorporates pixel-wise discriminators, importance-based pixel weighting, and an entropy-based alignment module to mitigate these problems. Experiments demonstrate that combining class reweighting with pixel-wise discrimination achieves a 2.4% improvement in mIoU over standard fine-tuning. While our approach improves performance on underrepresented classes, label alignment via fixed embeddings remains an open challenge, pointing to future opportunities for more adaptive alignment strategies.

Acknowledgment

This research was conducted under the university-corporate collaboration agreement between Kubota Corporation and the University of Tokyo.

References

[1] M. Cordts, M. Omran, S. Ramos, *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223.

[2] F. Yu, H. Chen, X. Wang, *et al.*, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2636–2645.

[3] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4990–4999.

[4] Y. Ganin, E. Ustinova, H. Ajakan, *et al.*, "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.

[5] A. Tavera, F. Cermelli, C. Masone, and B. Caputo, "Pixel-by-pixel cross-domain alignment for few-shot semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1626–1635.

[6] K. You, M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Universal domain adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2720–2729.

[7] T. Kalluri, G. Varma, M. Chandraker, and C. Jawahar, "Universal semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5259–5270.

[8] P. O. Bressan, J. M. Junior, J. A. Correa Martins, *et al.*, "Semantic segmentation with labeling uncertainty and class imbalance applied to vegetation mapping," *International Journal of Applied Earth Observation and Geoinformation*, vol. 108, p. 102 690, 2022, ISSN: 1569-8432. DOI: https://doi.org/10.1016/j.jag.2022.102690. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0303243422000162.

[9] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "Pidnet: A real-time semantic segmentation network inspired by pid controllers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 529–19 539.

[10] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016, pp. 761–769.

[11] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, 2021, pp. 8748–8763.