KDA: Knowledge Distillation Adversarial Framework With Vision Foundation Models for Landslide Segmentation

Shijie Wang, Lulin Li, Xuan Dong, Lei Shi, and Pin Tao

Abstract—**Landslides pose severe threats to infrastructure and safety, and their segmentation in remote sensing imagery remains challenging due to irregular boundaries, scale variation, and complex terrain. Traditional lightweight models often struggle to capture rich semantic features under these conditions. To address this, we leverage vision foundation models (VFMs) as teachers and propose a knowledge distillation adversarial (KDA) framework to transfer high-capacity knowledge into compact student models. Additionally, we introduce a dynamic cross-layer fusion (DCF) decoder to enhance global–local feature interaction. The experimental results demonstrate that, compared to the previous best-performing model SegNeXt [89.92% precision and 84.78% mean intersection over union (mIoU)], our method achieves a precision of 91.93% and mIoU of 86.53%, yielding improvements of 2.01% and 1.75%, respectively. Source code is available at https://github.com/PreWisdom/KDA**

Index Terms—**High-resolution remote sensing images, knowledge distillation (KD), landslide segmentation, semantic segmentation, vision foundation models (VFMs).**

I. Introduction

LANDSLIDE identification is crucial for disaster prevention and management. Accurate segmentation of landslide areas supports a timely response. Although deep learning has advanced segmentation performance, challenges remain due to complex spatial patterns and large-scale variations [1].

Vision foundation models (VFMs) [2], [3], [4] achieve strong performance but are hard to deploy in resource-limited devices due to their high computational cost.

Knowledge distillation (KD) addresses model compression by transferring knowledge from a large teacher to a compact student. However, traditional KD methods face key limitations. First, soft label alignment fails to fully transfer spatial knowledge, as downsampling in teacher models irreversibly removes fine-grained features (e.g., edges and textures). Second, large capacity gaps create representation bottlenecks—forcing low-dimensional student features to mimic high-dimensional teacher outputs via linear projection leads to feature loss from dimensional mismatch [5].

Adversarial learning improves generalization and robustness, helping models learn effective features and capture fine-grained semantics in complex scenes [6], [7]. Inspired by this, we propose the KD adversarial (KDA) framework to train a lightweight yet powerful student model that inherits VFM capabilities for refined segmentation. Our main contributions are as follows.

1) We introduce KDA to address a key limitation of traditional KD—student models often fail to capture fine-grained spatial semantics (Fig. 1) when the capacity gap is large.
2) We design a segmentation model combining a distilled backbone with a novel decoder. The decoder uses dynamic multiscale and cross-layer fusion to retain spatial details across scales while ensuring efficiency.
3) Extensive experiments show that our model outperforms SOTA methods, with a lightweight design suitable for resource-limited devices.

While KDA shows strong performance in our experiments, it has the following limitations.

1) *Sensitivity to Architecture:* Distillation becomes less effective when the teacher is a CNN and the student is a Transformer, suggesting that cross-architecture knowledge transfer needs improvement.
2) *Empirical Loss Weighting:* The KL divergence and mse loss weights (0.6/0.4) are manually tuned via grid search, lacking an adaptive weighting mechanism.

We aim to address these limitations in future work.

II. Related Work

A. Landslide Segmentation

Landslide segmentation enables pixel-level detection of slope failures via remote sensing, advancing geohazard analysis. Deep learning drives this progress by automatically

Received 1 April 2025; revised 7 June 2025 and 27 July 2025; accepted 6 August 2025. Date of publication 11 August 2025; date of current version 29 August 2025. *(Corresponding author: Pin Tao.)*

Shijie Wang and Lulin Li are with the School of Computer Technology and Applications, Qinghai University, Xining 810016, China (e-mail: mrggwsjj@gmail.com; lilulin188@gmail.com).

Xuan Dong is with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: dongxuan8811@bupt.edu.cn).

Lei Shi is with ACT&BDBC, Beihang University, Beijing 100191, China (e-mail: leishi@buaa.edu.cn).

Pin Tao is with the School of Computer Technology and Applications, Qinghai University, Xining 810016, China, also with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, and also with the Key Laboratory of Pervasive Computing, Ministry of Education, Beijing 100084, China (e-mail: taopin@tsinghua.edu.cn).

Digital Object Identifier 10.1109/LGRS.2025.3597685

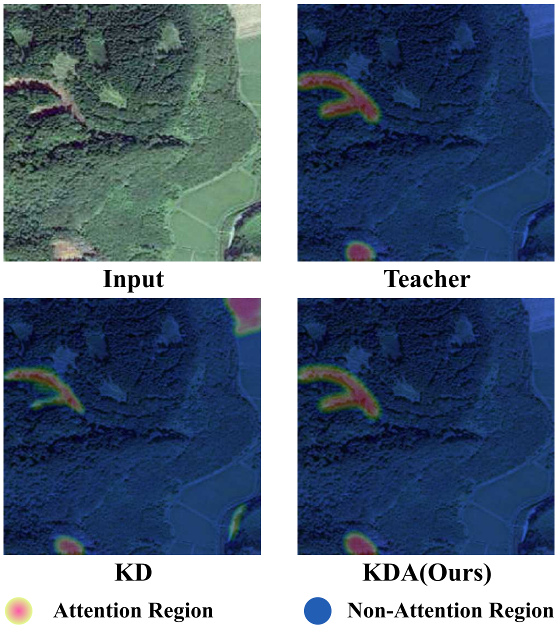

Fig. 1. Heatmap visualization. KD mislabels farmlands as landslides, and KDA better aligns with the teacher model.

learning features from Earth observation data [8], [9], [10], [11], [12], [13].

Initial breakthroughs enhanced local feature discrimination: Ji et al. [14] proposed 3-D spatial-channel attention; Soares et al. [15] improved U-Net with fused topographic-spectral features; Wang et al. [16] introduced boundary-aware Swin-Transformers; and Chandra et al. [17] explored channel-spatial attention synergies.

Current methods achieve SOTA in morphology tasks but struggle with cross-regional generalization (e.g., training on the Himalayas, deploying in the Andes) and spectral ambiguity (e.g., distinguishing landslides from quarries or roads).

B. Vision Foundation Models

VFMs like SAM [2] and DINO [3] have transformed computer vision with zero-shot capabilities. They handle morphological diversity and scene variations without relying on expensive labeled data, which is especially valuable.

Adapting VFMs to specialized tasks remains inefficient: full fine-tuning yields best results but is computationally costly [18]. Parameter-efficient methods offer alternatives—adapters [19] add lightweight modules (e.g., LandslideNet [20]), LoRA decomposes weight updates (e.g., SAMed [21]), and prompt learning tunes input tokens (e.g., RSPrompter [22]). They retain the full VFM at inference, limiting edge deployment.

C. Knowledge Distillation

KD [23] compresses models by transferring knowledge from a high-capacity teacher to a lightweight student. The student learns by minimizing divergence—typically KL divergence—between their output distributions, capturing the teacher's representations. However, traditional KD often fails to fully convey the rich information needed for accurate segmentation.

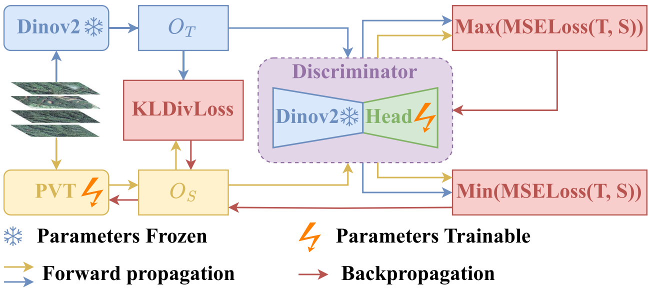

Fig. 2. Overview of our KDA framework. O_T and O_S are the teacher and student models' outputs. T and S are the discriminator's outputs after processing O_T and O_S. The discriminator improves by maximizing the T–S discrepancy, while the student model improves by minimizing it.

III. Methodology

A. KDA Framework

To address traditional KD's limitations in transferring representational capacity, we propose the KDA framework to effectively distill semantic knowledge from a high-capacity teacher to a lightweight student.

As shown in Fig. 2, the input data are processed simultaneously through two networks: a frozen pretrained teacher model (Dinov2) and a trainable student model (pyramid vision transformer, PVT [24]), producing feature representations O_T and O_S, respectively. The output distributions of the teacher and student are, then, aligned using KL divergence. Next, we introduce a discriminator (with a frozen Dinov2 backbone and a multilevel convolutional head) that receives outputs from both models and is tasked with distinguishing whether a given output originates from the teacher or the student.

The student model is trained with a loss combining KL divergence and mse loss from the discriminator. The discriminator uses mse loss to maximize the output gap between the teacher and student, while the student minimizes it, forming a minimax game. KL divergence aligns teacher and student output distributions, helping the student learn high-level semantics. MSE loss works at the pixel level, guiding the student to capture local structures like boundaries and details. This dual alignment enables the student to mimic the teacher's decisions while better recovering spatial details, improving segmentation performance.

Our framework naturally avoids two major GAN issues: mode collapse and training instability. Unlike GAN generators that produce low-diversity or noisy outputs early on, our student's distillation output is semantically rich, eliminating the need for high diversity or noise-like outputs and thus sidestepping these problems.

B. Model Architecture

In Fig. 3(a), our model integrates a distilled PVT backbone refined by KDA with a decoder. The backbone extracts multiscale semantic features rich in spatial and contextual details from remote sensing images. These features feed into the decoder, which uses dynamic multiscale selection and cross-layer fusion to progressively upsample and combine them, producing the final segmentation. This architecture combines a high-performance backbone with an innovative decoder, optimized for precise landslide boundary extraction.

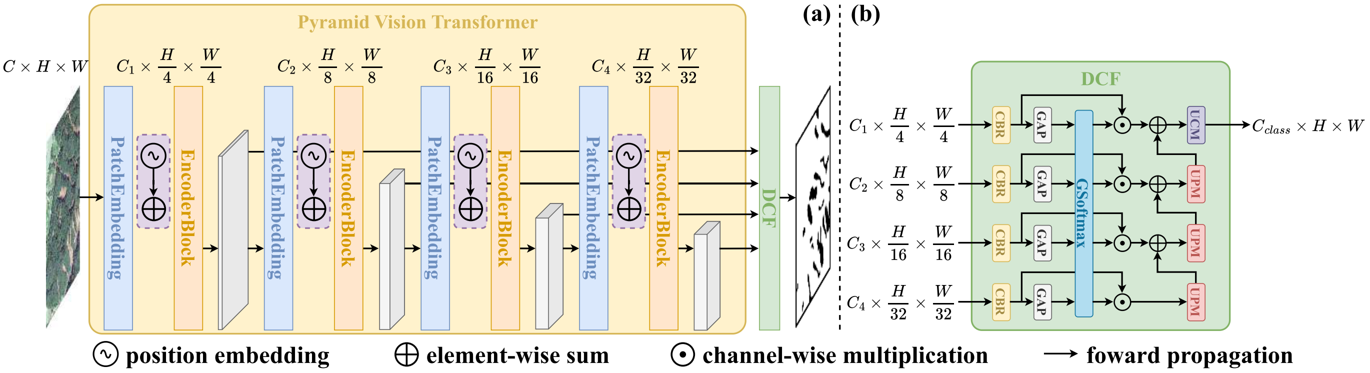
Pyramid Vision Transformer
(a)
(b)
C × H × W
$C_1 \times \frac{H}{4} \times \frac{W}{4}$
$C_2 \times \frac{H}{8} \times \frac{W}{8}$
$C_3 \times \frac{H}{16} \times \frac{W}{16}$
$C_4 \times \frac{H}{32} \times \frac{W}{32}$
PatchEmbedding
EncoderBlock
DCF
CBR
GAP
GSoftmax
UCM
UPM
$C_{class} \times H \times W$
position embedding
element-wise sum
channel-wise multiplication
foward propagation

Fig. 3. CBR: convolution BatchNorm ReLU module. GAP: global average pooling module. UPM: upsampling module. (a) Overall architecture: the input image passes through the distilled backbone (yellow, PVT-based), outputting multiscale features (4×, 8×, 16×, and 32×). These features enter the DCF decoder (green) to produce the segmentation mask. (b) DCF decoder: for each feature (F_i): 1) CBR module adjusts channels; 2) GAP and GSoftmax compute dynamic weights (A_i); 3) weights refine features ($\odot$ with F'_i); and 4) refined features (F^*_i) are progressively upsampled and fused across scales (cross-layer fusion) to combine coarse and fine details, outputting the result.

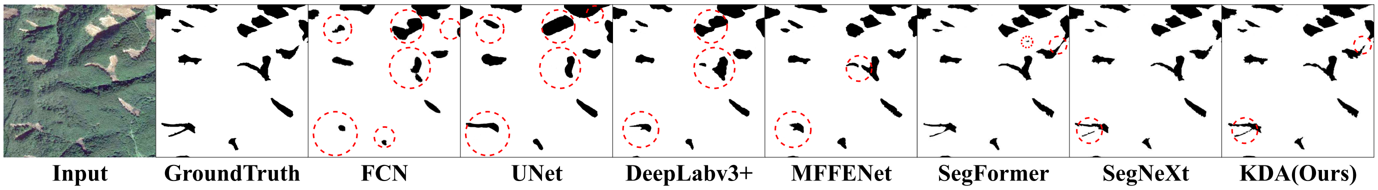

Fig. 4. This image compares our KDA method with mainstream landslide segmentation approaches (red circles highlight key differences). KDA achieves more accurate boundary localization and handles fine details like edges and small fragments better.

C. Dynamic Cross-Layer Fusion Decoder

In Fig. 3(b), our dynamic cross-layer fusion (DCF) decoder inputs four feature maps at 4×, 8×, 16×, and 32× downsampling. Each passes through a convolution-BatchNorm-ReLU (CBR) module to unify channel dimensions. Then, global average pooling (GAP) produces channelwise attention maps, refined by a shared-weight GSoftmax to generate dynamic channel weights. These weights multiply the pre-GAP features via skip connections, enhancing multiscale feature representation. The process is as follows:

$$\begin{aligned} F'_i &= \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{3\times3}(F_i))) \\ A_i &= \mathrm{GSoftmax}(\mathrm{GAP}(F'_i)) \\ F^*_i &= A_i \odot F'_i \end{aligned} \tag{1}$$

where F_i denotes the feature map extracted from the ith stage of the backbone, and F'_i represents the feature processed by the CBR module. A_i represents the channel attention weights of the ith stage, obtained through GAP and GSoftmax operations. F^*_i denotes the attention-refined feature. The operator $\odot$ signifies the channelwise Hadamard product.

The decoder progressively upsamples and fuses features from low to high resolution, combining coarse semantics with fine spatial details. An upsampling and contextual module (UCM), then, produces the final segmentation mask.

By dynamically computing channel attention, the decoder adaptively emphasizes important features, suppresses noise from low-value pixels, enhances multiscale integration, and reduces computational cost.

IV. Experiments

A. Experimental Settings

1) Dataset: We use the CAS Landslide dataset [25], which contains 20865 high-resolution (512 × 512) images from satellites and drones across nine regions. The dataset is split into training, validation, and test sets with an 8:1:1 ratio. It captures diverse environments, offering rich variability essential for training and evaluating landslide segmentation models.

2) Implementation Details: The experiments were conducted on a machine equipped with an NVIDIA 3090 GPU with 24 GB of memory, running Ubuntu 22.04. The software environment consisted of PyTorch 2.6.0, Python 3.11, and CUDA 12.4. To prevent overfitting and improve generalization, we applied a consistent data augmentation strategy across all experiments, which included random rotation, horizontal flipping, and vertical flipping.

In the first phase, it was trained with 200 epochs. The batch size was set to 8. The distillation temperature was set to 5. The discriminator's learning rate started at 5×10^{-6} and was gradually decreased to 5×10^{-9}. The student's learning rate started at 5×10^{-5} and was progressively reduced to 5×10^{-8}. The AdamW optimizer was used, and the discriminator's loss function L_D being

$$L_D = \max(L_{\mathrm{MSE}}(D_T, D_S)) \tag{2}$$

where D_T and D_S denote the discriminator's outputs when processing the features from the teacher and student models,

TABLE I
LOSS WEIGHT RATIO

Combination	mIoU↑	F1↑
0.5 × KL + 0.5 × MSE	85.97	90.61
0.6 × KL + **0.4** × MSE	**86.53**	**91.44**
0.7 × KL + 0.3 × MSE	86.11	91.25

TABLE II
COMPARISON TO STATE-OF-THE-ART MODELS ON CAS LANDSLIDE DATASET

Model	IoU↑	mIoU↑	OA↑	F1↑	Precision↑
FCN [26]	55.72	75.17	94.97	72.13	62.98
UNet [27]	51.18	72.62	94.41	70.60	61.80
DeepLabv3+ [28]	68.14	82.40	96.88	81.26	74.28
MFFENet [29]	67.99	82.32	96.86	81.15	74.14
SegFormer [30]	68.94	84.62	95.52	91.20	90.27
SegNeXt [31]	69.05	84.78	97.24	90.80	89.92
KDA(Ours)	**70.76**	**86.53**	**98.18**	**91.44**	**91.93**

respectively. The student's loss function L_S defined as

$$L_S = 0.4 \times \min(L_{\text{MSE}}(D_T, D_S)) + 0.6 \times L_{\text{KLDiv}}(O_T, O_S) \quad (3)$$

where O_T and O_S denote the outputs from the teacher and student models, respectively.

We use a weighted loss of 0.6× KL +0.4× mse to balance semantic alignment and output diversity. A higher KL weight stabilizes semantic learning, while moderate mse prevents overfitting, enhancing generalization. To validate this choice, we conducted an ablation study comparing three loss combinations on the validation set.

Table I shows that the 0.6:0.4 ratio achieved the best performance. A higher mse weight (>0.5) led to training instability or mode collapse (common in adversarial settings), while a lower weight weakened adversarial guidance. Thus, 0.6:0.4 offers a practical, empirically validated balance between performance and stability.

For the second phase, it was also trained with 200 epochs, the batch size was set to 16. The backbone's learning rate began at 5×10^{-6} and was reduced to 5×10^{-9}, while the decoder's learning rate started at 5×10^{-4} and gradually decreased to 5×10^{-6}. The AdamW optimizer was again used, with the segmentation model's loss function L_M being

$$L_M = 0.4 \times L_{\text{CE}}(\text{GT}, O) + 0.6 \times L_{\text{Focal}}(\text{GT}, O) \quad (4)$$

where $\text{GT} \in \{0, 1\}^{H \times W}$ denotes the binary ground-truth mask and $O \in \mathbb{R}^{H \times W}$ represents the output from the segmentation model, with H and W indicating spatial dimensions. The above loss weight ratio was determined via grid search to achieve optimal performance.

3) Evaluation Metrics: We used precision, overall accuracy (OA), $F1$-score ($F1$), intersection over union (IoU), and mean IoU (mIoU) as accuracy evaluation metrics. These metrics are commonly employed to assess the performance of segmentation models.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT TEACHER MODELS AND KD METHODS

Method	Teacher	IoU↑	mIoU↑	OA↑	F1↑
KD	Dinov2-B	68.82	84.67	96.64	88.93
	Dinov2-L	69.11	84.86	96.32	88.62
	Dinov2-G	69.25	85.17	96.25	89.26
	ConvNeXt-B	68.61	84.30	96.20	88.73
	ConvNeXt-L	68.79	84.59	95.83	88.51
	ConvNeXt-XL	69.20	84.96	96.72	89.17
KDA(Ours)	Dinov2-B	69.18	84.91	96.54	89.08
	Dinov2-L	**70.76**	**86.53**	**98.18**	**91.44**
	Dinov2-G	69.34	85.29	97.07	90.23
	ConvNeXt-B	68.44	84.16	96.78	88.72
	ConvNeXt-L	68.96	84.74	97.08	90.12
	ConvNeXt-XL	69.24	85.13	97.14	89.92

B. Comparison to State-of-the-Art Methods

As shown in Table II, the KDA method significantly outperformed other models [26], [27], [28], [29], [30], [31] on the CAS dataset, achieving SOTA performance. The intuitive comparisons of KDA with other methods are highlighted in Fig. 4. Compared to the previous best-performing model, SegNeXt, KDA improved precision by 2.01%, IoU by 1.71%, and mIoU by 1.75%, demonstrating its effectiveness.

C. Ablation Studies

1) Distillation Method: First, to investigate the impact of KD methods on segmentation performance, we employ two distillation strategies—traditional KD and our KDA. Using a unified student model (PVT) with selected teacher architectures, the distillation process is systematically evaluated through IoU, mIoU, OA, and $F1$ metrics. This controlled comparison allows us to isolate the effects of different knowledge transfer methodologies.

As shown in Table III, the KDA method outperforms the KD method in model performance across the board, achieving significant improvements in key metrics. For instance, with Dinov2-L as the teacher model, IoU increased by 1.65%, mIoU improved by 1.67%, OA increased by 1.86%, and $F1$ increased by 2.82%. In addition, KDA also shows a stable performance improvement with other teacher models (e.g., ConvNeXt-L), demonstrating its generalization ability.

2) Teacher Model: Second, to examine the scaling effects of teacher models, we conduct comparative experiments across six model scales spanning both Dinov2 and ConvNeXt families. The consistent use of PVT as a student model ensures fair evaluation of teacher capacity influences.

As shown in Table III, increasing the size of the teacher model does not lead to a significant performance boost. For example, in the KDA method, the IoU only improved by 0.28% when using ConvNeXt-XL instead of ConvNeXt-L. Moreover, we observe that Transformer-based teacher models Dinov2 yield better results than CNN-based teacher models ConvNeXt, likely due to architectural compatibility with the Transformer-based student. When performance is comparable, smaller and structurally aligned teacher models are preferable, as they maintain effectiveness while reducing distillation costs.

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT DECODERS. P AND F REPRESENT THE PARAMETERS (M) AND FLOPS (G), RESPECTIVELY

Decoder	IoU↑	mIoU↑	OA↑	F1↑	P↓	F↓
DCF(Ours)	**70.76**	**86.53**	**98.18**	**91.44**	**0.27**	**22.26**
UNet-Head [26]	55.87	75.04	72.34	69.92	12.16	35.52
SegFormer-Head [30]	68.94	84.62	95.52	91.20	0.76	24.08
SegNeXt-Head [31]	65.24	81.63	93.72	80.87	3.87	28.09
UperNet-Head [32]	66.74	80.63	94.57	85.79	0.97	23.04

3) Decoder Architecture: Finally, to validate the effectiveness of our proposed DCF decoder, we compared it with the decoders of popular segmentation models [27], [30], [31], [32] under the same experimental setup. Using the PVT distilled with the KDA method as the backbone.

As shown in Table IV, DCF achieves the highest scores with the lowest computational cost. It is over 2× lighter in parameters than UperNet-Head [30] (0.97M) and SegFormer-Head [28] (0.76M), and vastly more efficient than UNet-Head [24] (12.16M/35.52G). This establishes DCF as the new SOTA in the accuracy-efficiency tradeoff.

V. CONCLUSION

We propose the KDA framework to enhance knowledge transfer through adversarial training. KDA achieves new SOTA performance with 86.53% mIoU and 91.93% precision, surpassing SegNeXt by 1.75% and 2.01%, respectively. With only 0.27M parameters, the model is lightweight and suitable for deployment on resource-constrained devices.

Despite its strong performance, KDA has two key limitations. First, its effectiveness decreases when the teacher and student use different architectures (e.g., CNN versus Transformer), highlighting the need for improved cross-architecture transfer. Second, the loss weight ratio was determined through grid search and lacks adaptivity.

To address these issues, future work will focus on incorporating intermediate-layer feature alignment to bridge architectural gaps and developing a dynamic loss weighting strategy to adaptively balance semantic alignment and spatial detailed learning during training.

REFERENCES

[1] A. Mohan, A. K. Singh, B. Kumar, and R. Dwivedi, "Review on remote sensing methods for landslide detection using machine and deep learning," *Trans. Emerg. Telecommun. Technol.*, vol. 32, no. 7, p. 3998, Jul. 2021.

[2] A. M. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2023, pp. 4015–4026.

[3] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9660.

[4] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," 2023, *arXiv:2304.07193*.

[5] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.

[6] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[8] L. Li, B. Chen, X. Zou, J. Xing, and P. Tao, "UV-mamba: A DCN-enhanced state space model for urban village boundary identification in high-resolution remote sensing images," 2024, *arXiv:2409.03431*.

[9] J. Li, X. Zou, S. Wang, B. Chen, J. Xing, and P. Tao, "A parallel attention network for cattle face recognition," 2024, *arXiv:2403.19980*.

[10] B. Chen et al., "LEFormer: A hybrid CNN-transformer architecture for accurate lake extraction from remote sensing imagery," 2023, *arXiv:2308.04397*.

[11] B. Chen, X. Zou, K. Li, Y. Zhang, J. Xing, and P. Tao, "High-fidelity lake extraction via two-stage prompt enhancement: Establishing a novel baseline and benchmark," 2023, *arXiv:2308.08443*.

[12] Z. Zheng, L. Lv, J. He, and L. Zhang, "UniRS: Toward unified multitask fine-tuning for remote sensing foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5622413, doi: 10.1109/TGRS.2025.3564609.

[13] L. Lv and L. Zhang, "Advancing data-efficient exploitation for semi-supervised remote sensing images semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5619213, doi: 10.1109/TGRS.2024.3388199.

[14] S. Ji, D. Yu, C. Shen, W. Li, and Q. Xu, "Landslide detection from an open satellite imagery and digital elevation model dataset using attention boosted convolutional neural networks," *Landslides*, vol. 17, no. 6, pp. 1337–1352, Jun. 2020.

[15] L. P. Soares, H. C. Dias, and C. H. Grohmann, "Landslide segmentation with U-net: Evaluating different sampling methods and patch sizes," 2020, *arXiv:2007.06672*.

[16] X. Wang et al., "Refined intelligent landslide identification based on multi-source information fusion," *Remote Sens.*, vol. 16, no. 17, p. 3119, Aug. 2024.

[17] N. Chandra, H. Vaidya, S. Sawant, and S. R. Meena, "A novel attention-based generalized efficient layer aggregation network for landslide detection from satellite data in the higher himalayas, Nepal," *Remote Sens.*, vol. 16, no. 14, p. 2598, Jul. 2024.

[18] X. Sun et al., "RingMo: A remote sensing foundation model with masked image modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2022, Art. no. 5612822.

[19] N. Houlsby et al., "Parameter-efficient transfer learning for NLP," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2790–2799.

[20] J. Yu et al., "Landslidenet: Adaptive vision foundation model for landslide detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2024, pp. 7282–7285.

[21] K. Zhang and D. Liu, "Customized segment anything model for medical image segmentation," 2023, *arXiv:2304.13785*.

[22] K. Chen et al., "RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4701117.

[23] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[24] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[25] Y. Xu, C. Ouyang, Q. Xu, D. Wang, B. Zhao, and Y. Luo, "CAS landslide dataset: A large-scale and multisensor dataset for deep learning-based landslide detection," *Scientific Data*, vol. 11, no. 1, p. 12, Jan. 2024.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[28] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[29] Q. Xu, C. Ouyang, T. Jiang, X. Yuan, X. Fan, and D. Cheng, "MFFENet and ADANet: A robust deep transfer learning method and its application in high precision and fast cross-scene recognition of earthquake-induced landslides," *Landslides*, vol. 19, no. 7, pp. 1617–1647, Jul. 2022.

[30] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Sys. (NIPS)*, vol. 34, Dec. 2021, pp. 12077–12090.

[31] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M. Cheng, and S. Hu, "SegNeXt: Rethinking convolutional attention design for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 1140–1156.

[32] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 418–434.