electronics

MDPI

Article

PLAID: Pyramid-Level Adaptive Imbalance-Aware Diversification for Source-Free Domain Adaptation in Power-Grid Temporal Action Detection

Lingwen Meng [1], Siwu Yu [1], Mingyong Xin [1,*] and Xinshan Zhu [2]

[1] Electric Power Research Institute of Guizhou Power Grid Co., Ltd., Guiyang 550002, China; menglw@gz.csg.cn (L.M.); yusw@gz.csg.cn (S.Y.)
[2] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China; xszhu@tju.edu.cn
* Correspondence: xinmy@csg.cn

Abstract

Temporal action detection (TAD) localizes safety-critical procedures in untrimmed power-grid operation videos. Cross-site changes in action duration and density create level-dependent feature collapse in temporal pyramids and noisy, incomplete pseudo-labels under source-free domain adaptation. We propose Pyramid-Level Adaptive Imbalance-Aware Diversification (PLAID), which augments Mean-Teacher self-training with pyramid-level representation regularization. Duration-Aware Sampling (DAS) first constructs token matrices at duration-matched pyramid levels. Cross-Level Borrowing (CLB) then supplies temporally aligned neighboring tokens to levels with limited pseudo-segment support, stabilizing the statistics used by subsequent constraints. Channel variance and decorrelation losses preserve feature diversity, while Rank-Gap Reweighting (RGR) scales these constraints according to each level's effective-rank decline from its running maximum. Together, these label-free components reduce reliance on imperfect pseudo-labels without source videos or target annotations. On four controlled duration and density shifts in the Power Grid Operations benchmark, PLAID improves source-only ActionFormer by 2.5–6.0 mean average precision (mAP) points and outperforms Vanilla Mean-Teacher in every setting. With TemporalMaxer, it also surpasses both baselines at every evaluated temporal intersection-over-union (tIoU) threshold in two representative shifts. The results support pyramid-aware representation recovery as a practical strategy for adapting fixed-camera power-grid detectors across sites when operational tempo and action occupancy vary and source recordings cannot be shared.

Keywords: temporal action detection; source-free domain adaptation; power-grid operation monitoring; representation regularization; temporal feature pyramid

Check for updates

Academic Editor: John A. Kalomiros

Received: 15 October 2026
Revised: 5 November 2026
Accepted: 8 November 2026
Published: 18 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Power-grid operation monitoring records each on-site maintenance procedure as an untrimmed video. Temporal Action Detection (TAD) localizes and classifies safety-critical procedures such as grounding-rod handling and verification, thereby enabling automatic identification of non-compliant operations. A common source of temporal variation is grounding-rod operation: differences in workspace constraints and operator tempo can lengthen or shorten individual rod-handling actions, while site-specific preparation, inspection, and waiting intervals change the fraction of the recording occupied by active operations. Cross-site deployment can therefore shift both absolute action duration and

relative action occupancy, the latter referred to as action density. These shifts perturb the regression and classification priors, respectively, and redistribute action instances across the levels of multi-scale temporal detectors such as ActionFormer [1]. To isolate their effects, we evaluate four directed source-to-target transfers in the Power Grid Operations (PGO) benchmark: Long-to-Short (PGO-L2S), Short-to-Long (PGO-S2L), Sparse-to-Dense (PGO-S2D), and Dense-to-Sparse (PGO-D2S). PGO-L2S and PGO-S2L vary absolute action duration, whereas PGO-S2D and PGO-D2S vary relative action occupancy. The closest generalizable or domain-adaptive temporal action localization methods either study action-scale transfer under weak supervision [2] or require source-domain data during adaptation [3]; they do not jointly address duration and density shifts under source-free deployment.

Source-free adaptation is particularly important in the power-grid setting because site videos may contain operationally sensitive infrastructure and are difficult to retain or transfer across regional utilities. Source-Free Domain Adaptation (SFDA) [4] removes this requirement by adapting from a source-pretrained model and unlabeled target videos. However, when temporal statistics differ substantially across sites, the source detector provides unreliable target-domain predictions, making adaptation vulnerable to noisy pseudo-labels and uneven degradation across temporal-pyramid levels.

Representation degradation constitutes another challenge. As shown in Figure 1A,B, applying the source-only model to target videos lowers effective rank at all six levels in both duration-shift directions. Figure 1C further shows that Mean-Teacher self-training changes this profile unevenly in PGO-S2D, with particularly severe rank loss at the first four levels. This level-dependent behavior motivates pyramid-level representation regularization that recovers lost feature diversity rather than applying the same treatment to every level.

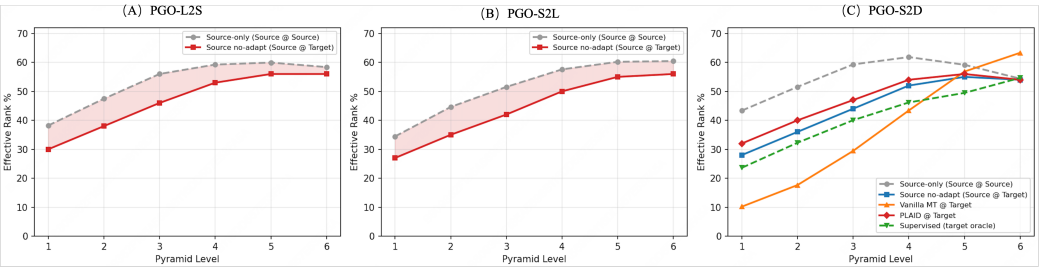


Figure 1. Effective-rank profiles across six temporal-pyramid levels for the Power Grid Operations (PGO) benchmark. L2S, S2L, and S2D denote long-to-short, short-to-long, and sparse-to-dense transfers, respectively. Pyramid-Level Adaptive Imbalance-Aware Diversification (PLAID) is our method. Levels 1–6 correspond to $l = 0, \ldots, 5$. (**A**,**B**) The source-only model exhibits rank reductions after transfer in PGO-L2S and PGO-S2L (Source @ Target versus Source @ Source). (**C**) In PGO-S2D, PLAID mitigates the low-level rank collapse of Vanilla Mean-Teacher and moves the target-domain rank closer to the source reference at Levels 1–5. Vanilla Mean-Teacher attains the highest rank at Level 6, illustrating that PLAID preserves non-degenerate, level-specific feature geometry rather than maximizing rank.

The standard SFDA pipeline relies on Mean-Teacher self-training [5]. An exponential moving average (EMA) teacher generates pseudo-labels on target videos, and the student is optimized against those labels. Pseudo-label quality becomes the principal limitation when the source model is unreliable on the target domain because incorrect temporal boundaries, class labels, or both can be reinforced during adaptation. On our power-grid benchmark, source-only performance ranges from 23% to 32% target mean average precision (mAP) under substantial domain shift. Figure 2a quantifies the resulting precision–recall trade-off in PGO-S2D: at the deployed threshold $\tau = 0.5$, pseudo-label precision is 29.9%, whereas the highest tested F1 is only 0.240 at $\tau = 0.4$. Figure 2b further shows that false positives dominate most confidence bins, although their relative prevalence decreases at

high confidence. Confidence thresholding alone therefore cannot provide supervision that is simultaneously reliable and sufficiently complete in this setting.

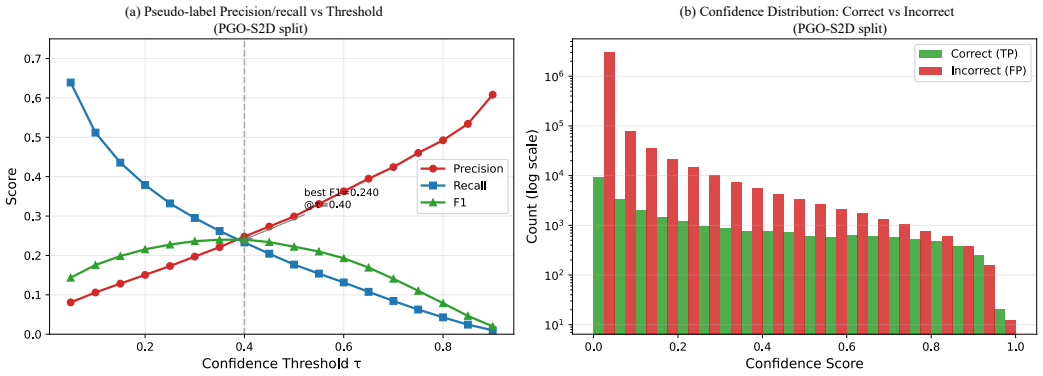


Figure 2. Pseudo-label confidence analysis in PGO-S2D under the temporal intersection-over-union (tIoU) ≥ 0.5 matching criterion. PGO-S2D denotes the sparse-to-dense transfer on the Power Grid Operations benchmark. (**a**) Precision, recall, and F1 across confidence thresholds. Increasing τ improves precision but reduces recall; the dashed line marks the highest tested F1 of 0.240 at $\tau = 0.4$, while the deployed threshold $\tau = 0.5$ yields 29.9% precision. (**b**) Counts of true-positive (TP) and false-positive (FP) predictions across confidence bins on a logarithmic scale. False positives dominate most bins, although the gap narrows at high confidence. Confidence filtering therefore cannot provide supervision that is both reliable and sufficiently complete in this setting.

Pyramid-Level Adaptive Imbalance-Aware Diversification (PLAID) responds to these observations by complementing confidence-based self-training with pyramid-level representation regularization. In PGO-S2D, Figure 1C shows that PLAID moves the target-domain rank toward the Source @ Source reference at the first five levels and mitigates the pronounced Mean-Teacher rank loss at Levels 1–4. At Level 6, however, Vanilla Mean-Teacher exceeds the source-domain reference, PLAID, and the target-supervised oracle. The figure therefore motivates preserving non-degenerate, level-specific feature geometry rather than maximizing rank or reproducing a fixed rank profile. Channel variance and decorrelation provide the basic recovery force by preventing inactive and redundant feature directions. Rank-Gap Reweighting (RGR) distributes this force across levels according to their rank evolution during adaptation, and Cross-Level Borrowing (CLB) increases sample support for the statistics used by the reweighting mechanism. Section 3 specifies the four components.

The contributions are summarized as follows.

- PLAID introduces pyramid-level adaptive representation diversification for source-free domain adaptation in TAD (SFDA-TAD). Duration-aware token assignment, cross-level statistical support, and history-relative rank-gap weighting jointly determine the feature population, statistics, and regularization strength at each temporal-pyramid level. This design addresses the interaction between uneven representation degradation and unequal statistical support across levels within source-free self-training.
- Evaluation on controlled duration and action-density splits of the Power Grid Operations benchmark shows gains of 2.5 to 6.0 mAP over source-only and improvements over Vanilla Mean-Teacher on all four splits. In the ablation, the adaptive gain appears only under the joint use of RGR and CLB, a pattern consistent with CLB supplying the sample support required by RGR.
- Cross-architecture evaluation confirms that PLAID remains effective with TemporalMaxer, exceeding both source-only and Vanilla Mean-Teacher at every tested threshold

in the long-to-short and dense-to-sparse settings. This threshold-wise consistency shows that the gain is not confined to a single localization tolerance.

2. Related Work

2.1. Temporal Action Detection for Power-Grid Videos

Video analysis supports automatic safety monitoring in power-grid operations. Existing studies primarily formulate operator monitoring as behavior or violation classification. They address cross-domain few-shot transfer [6], hard-example learning [7], continual recognition under changing operating conditions [8], and real-time object–graph modeling [9]. Power-industry studies further examine skeleton-based video recognition for live-line work [10], pose-based personnel classification [11], and continual visual learning under distribution change [12]. Other components of operational monitoring include image-based dynamic risk identification [13], three-dimensional reconstruction of complex work scenes [14], safety-zone generation [15], operational-feature recognition [16], and pose-aware cleaning of safety-monitoring images [17]. These methods establish the practical value of power-grid video analysis, but do not address temporal localization in untrimmed videos under source-free cross-site deployment. TAD is therefore required to identify both the category and temporal boundaries of complete operational procedures.

TAD has progressed from proposal and boundary modeling [18,19] to direct temporal modeling [20–22] and multi-scale pyramid detectors such as ActionFormer [1] and TemporalMaxer [23]. Recent work further improves boundary refinement [24], cross-dataset generalization [2], and long-range temporal modeling [25]. TAD detectors commonly operate on pre-extracted features from Inflated 3D ConvNets (I3D) [26], with each pyramid level responsible for a different action-duration range. This scale structure is particularly consequential in power-grid deployment: cross-site duration and density shifts can affect pyramid levels unevenly, whereas existing TAD methods do not explicitly recover level-dependent representation degradation without source videos.

2.2. Source-Free Domain Adaptation

SHOT [4] established information maximization as a core objective for SFDA. Mean-Teacher [5] provides a complementary mechanism through exponential moving average targets and stable pseudo-labeling. Recent work directly addresses pseudo-label ambiguity [27], and source-free adaptation has also been studied for segmentation [28]. SADA studies unsupervised domain adaptation for temporal action localization, but requires source-domain data during adaptation [3]. Recent SFDA studies address unreliable source models, target clustering, prompt-guided alignment, and source-free video or time-series adaptation [29–33]. Related test-time adaptation work investigates online expectation-maximization and adaptive quantile recalibration [34,35]. These works do not directly address source-free temporal detection, but they reinforce the importance of stabilizing target-only adaptation.

Directly transferring these SFDA paradigms to TAD leaves two task-specific gaps. First, a temporal pseudo-label contains both an action category and temporal boundaries, so confidence filtering trades label reliability against supervision coverage. Second, TAD predictions are produced by a multi-level temporal pyramid whose levels specialize in different action-duration ranges, making representation degradation inherently level dependent. Generic image-based SFDA objectives do not explicitly account for this combination of temporal-boundary uncertainty and duration-specific level degradation. These gaps motivate an SFDA-TAD approach that complements pseudo-label self-training with pyramid-level representation recovery.

2.3. Representation Regularization

Representation regularization is commonly used to prevent collapse in self-supervised learning. SimCLR [36] and MoCo [37] established contrastive representation learning, while VICReg [38] combines variance, invariance, and covariance regularization. Recent work further studies covariance-based feature-space regularization [39], debiased representation learning [40], and extensions of VICReg to kernel feature spaces [41].

PLAID formulates representation diversification at the temporal-pyramid level, with duration-conditioned token populations and adaptive regularization strengths. The variance term discourages near-constant channel responses by penalizing standard deviations below a floor, while the decorrelation term penalizes off-diagonal channel covariance to reduce redundancy. PLAID extends these established constraints through duration-conditioned token matrices, cross-level statistical support, and rank-dependent regularization strength. DAS preserves the detector's duration-based feature decomposition. CLB augments the level-wise token matrices with temporally corresponding features from adjacent levels, while RGR computes bounded weights from each level's decline in normalized effective rank [42] relative to its own history. These operations jointly specify where the channel constraints act, which samples define their statistics, and how strongly each level is regularized.

Mean-Teacher supplies the self-training framework. The methodological contribution of PLAID lies in linking duration-aware sampling, cross-level statistical support, and rank-based allocation in a single level-adaptive objective. In this formulation, the allocation made by RGR depends on the token statistics constructed through DAS and CLB. This dependency distinguishes the objective from pooled feature diversification or uniformly weighted channel regularization. The ablation study in Section 4 evaluates the resulting coupling.

3. Method

The method design follows the two empirical observations established in Figures 1 and 2. First, source-to-target transfer reduces effective rank unevenly across pyramid levels, motivating level-specific recovery of useful feature diversity rather than uniform regularization or numerical rank maximization. Second, the PGO-S2D analysis in Figure 2 shows that confidence filtering cannot provide pseudo-labels that are both reliable and sufficiently complete: higher thresholds reduce false positives but discard many correct predictions, while the deployed threshold retains substantial noise. PLAID addresses the interaction of these problems by complementing the pseudo-label detection loss with a pyramid-level representation objective that requires no target annotations. It does not modify or denoise the selected pseudo-labels. Instead, channel variance and decorrelation directly counter representation collapse, Rank-Gap Reweighting (RGR) allocates stronger recovery to levels with greater rank decline, and Cross-Level Borrowing (CLB) stabilizes the statistics used at under-sampled levels. Confidence remains responsible for pseudo-label selection, while effective rank provides a complementary signal for allocating recovery strength.

3.1. Source-Free Mean-Teacher Framework

Let f_{θ_0} denote a temporal action detector with source-pretrained parameter vector θ_0. During adaptation, the source videos and annotations are unavailable, and only θ_0 and an unlabeled target set $\mathcal{D}_t = \{x_i\}_{i=1}^{N}$ are retained, where N is the number of target videos and $i \in \{1, \ldots, N\}$ is the video index. The objective is to adapt the detector to $\mathcal{D}_t$ without accessing source-domain samples. A Mean-Teacher framework maintains a student detector f_θ with parameters θ and an EMA teacher detector $f_{\bar{\theta}}$ with parameters $\bar{\theta}$, both initialized from θ_0. Given teacher momentum $\mu \in [0, 1)$, the teacher is updated after each student update by

$$\bar{\theta} \leftarrow \mu\bar{\theta} + (1 - \mu)\theta. \tag{1}$$

The EMA teacher generates pseudo-labels during adaptation and is used for evaluation because its parameters provide a temporally smoothed estimate of the student trajectory.

For a target video x_i, the teacher predicts temporal segments with class scores. Predictions whose maximum class score exceeds the confidence threshold τ are retained, Gaussian soft non-maximum suppression (Soft-NMS) is applied, and at most $N_{\max}$ predictions are preserved, where $N_{\max}$ is the maximum number of pseudo-segments retained per video. This process yields the pseudo-label set

$$\hat{y}_i = \{(\hat{s}_j, \hat{e}_j, \hat{c}_j)\}_{j=1}^{N_i}, \tag{2}$$

where x_i is the i-th target video, $0 \le N_i \le N_{\max}$ is the number of retained pseudo-segments for x_i, and $\hat{s}_j$, $\hat{e}_j$, and $\hat{c}_j$ are the predicted start time, end time, and action class of the j-th pseudo-segment, respectively. The student is optimized using the detector's standard classification and temporal-regression loss $\mathcal{L}_{\text{det}}$. Videos with no retained pseudo-segments do not contribute to $\mathcal{L}_{\text{det}}$.

For each input, the detector produces temporal-pyramid feature maps $\{F_l\}_{l=0}^{L-1}$, where L is the number of pyramid levels and $l \in \{0, \dots, L-1\}$ is the level index. Each map satisfies $F_l \in \mathbb{R}^{T_l \times C}$, where T_l is the number of temporal positions at level l and C is the shared feature-channel dimension. The input or mini-batch index of F_l is omitted when clear. PLAID operates on these pyramid features while preserving the Mean-Teacher pseudo-labeling and detection pipeline. The following sections describe how per-level token statistics are constructed and regularized.

3.2. PLAID Overview

PLAID augments Mean-Teacher self-training with pyramid-level representation regularization. As illustrated in Figure 3, the teacher first generates pseudo-segments for the standard detection loss. In parallel, Duration-Aware Sampling (DAS) assigns pseudo-segments to their duration-matched pyramid levels and constructs level-specific token matrices. Cross-Level Borrowing (CLB) then augments under-sampled levels with temporally aligned tokens from adjacent levels. Channel variance and decorrelation losses promote feature diversity on the resulting matrices, while Rank-Gap Reweighting (RGR) adaptively scales these losses according to the effective-rank decline observed at each level. PLAID therefore complements pseudo-label supervision without modifying the detector head or pseudo-label generation procedure. The following sections describe its four components in order: DAS, CLB, channel regularization, and RGR.

3.3. Duration-Aware Sampling

A representation regularizer requires statistics that correspond to the population processed by each pyramid level. Pooling foreground tokens from all pseudo-segments would conflate short and long actions assigned to different levels. The resulting variance and covariance would describe a cross-scale mixture rather than any individual level, and degradation at one level could be obscured by features from another. The sampling procedure must therefore preserve the detector's scale decomposition before regularization is applied.

Duration-Aware Sampling (DAS) preserves this decomposition by routing each pseudo-segment to the pyramid level whose predefined duration range contains its predicted duration, computed as the difference between its predicted end and start times. Let $\mathcal{S}_l$ collect the pseudo-segments dispatched to level l. DAS then draws n_{fg} foreground tokens from each pseudo-segment interior and n_{bg} background tokens from non-foreground positions on F_l, forming

$$Z_l = \left[F_l[\mathcal{I}_l^{\text{fg}}];\ F_l[\mathcal{I}_l^{\text{bg}}] \right] \in \mathbb{R}^{N_l \times C}, \quad N_l = n_{\text{fg}} \cdot |\mathcal{S}_l| + n_{\text{bg}}, \tag{3}$$

where $|\mathcal{S}_l|$ is the number of segments in $\mathcal{S}_l$, $\mathcal{I}_l^{\mathrm{fg}}$, and $\mathcal{I}_l^{\mathrm{bg}}$ are the sampled foreground and background time-index sets, and $F_l[\cdot]$ indexes rows of the level-l feature map. The semicolon denotes row-wise concatenation, and N_l is the resulting number of rows in Z_l. The fixed sampling budgets are $n_{\mathrm{fg}} = 8$ foreground tokens per pseudo-segment and $n_{\mathrm{bg}} = 32$ background tokens per level. This construction ensures that the downstream variance, decorrelation, and rank statistics are estimated from the tokens processed by the corresponding level rather than from a cross-scale mixture. The inclusion of background tokens ensures that these statistics cover both action and non-action regions instead of being dominated by either population.

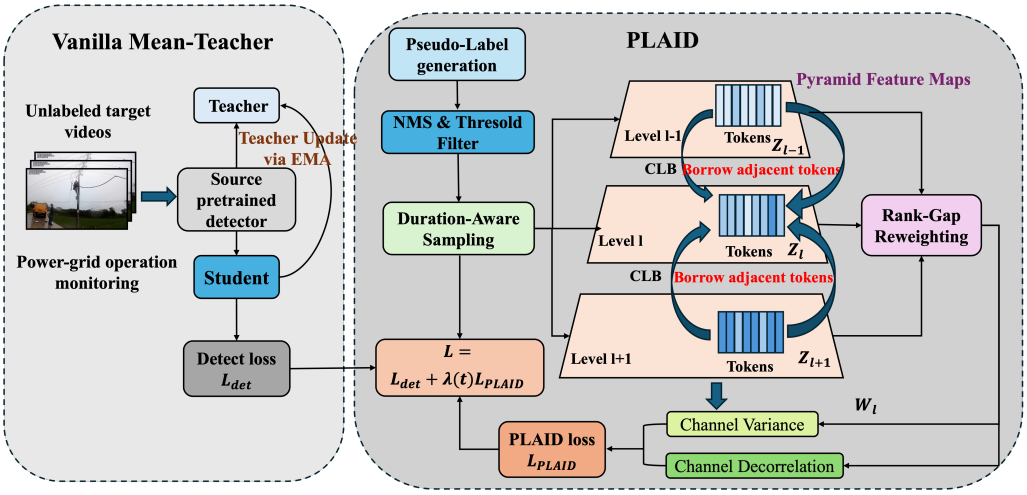


Figure 3. Overview of PLAID. Mean-Teacher features support the standard detection loss and pyramid-level regularization. Duration-Aware Sampling (DAS) constructs duration-matched, level-specific token matrices, which Cross-Level Borrowing (CLB) augments with stride-aligned tokens from adjacent levels. Variance and decorrelation losses preserve feature diversity, while Rank-Gap Reweighting (RGR) scales them according to each level's effective-rank gap. The total objective combines the detection loss with warmup-weighted PLAID regularization. EMA denotes exponential moving average, and NMS denotes non-maximum suppression.

3.4. Cross-Level Borrowing

Limited pseudo-segment support can make the rank and covariance statistics of an individual pyramid level sensitive to the sampled tokens. CLB supplements the current-level matrix with tokens from the immediately adjacent levels. This locality limits the scale discrepancy between the host and borrowed representations, but does not make the levels semantically interchangeable. The design rationale is to obtain additional temporal-context support for statistical regularization while retaining the detector's duration-based decomposition.

For each pyramid level l, CLB samples n_{borrow} tokens in total from its immediately adjacent levels ($l-1$ and $l+1$, when available) and maps their indices to the timeline of level l according to the relative temporal strides. Let $Z_l^{\mathrm{nbr}} \in \mathbb{R}^{n_{\mathrm{borrow}} \times C}$ contain the resulting tokens. After standardizing the borrowed tokens to the channel statistics of the current-level matrix Z_l, CLB forms

$$\tilde{Z}_l = \left[Z_l;\ \mathrm{standardize}\left(Z_l^{\mathrm{nbr}}; Z_l\right) \right] \in \mathbb{R}^{(N_l + n_{\mathrm{borrow}}) \times C}. \tag{4}$$

Here, N_l is the number of current-level tokens defined in Equation (3), and the semicolon denotes row-wise concatenation. The borrowing budget is n_{borrow} in total across the available immediate neighbors, not a separate budget for each neighbor. Temporal correspondence is defined on the common video timeline through the levels' relative temporal strides. It identifies corresponding temporal locations, rather than requiring identical receptive fields or action semantics.

Borrowed tokens are standardized independently in each feature channel and then rescaled to the channel-wise mean and standard deviation of Z_l. This operation reduces marginal scale differences; matching the first two channel moments does not align the complete joint distribution or establish class consistency. Borrowed tokens supplement the matrices used by the rank, variance, and decorrelation statistics. They do not replace the host feature map used by the detection head, change duration-based level assignment, or create additional pseudo-labeled detection instances.

When CLB is disabled, $n_{\text{borrow}} = 0$ and $\tilde{Z}_l = Z_l$. CLB uses adjacent-scale temporal context to supplement the foreground–background statistics used by the channel constraints, without transferring class labels between levels. Section 4 evaluates its interaction with RGR.

3.5. Channel Variance and Decorrelation

Noisy self-training can reduce the dimensionality of per-level features through two mechanisms. Individual channels may converge toward constant responses, which produces variance collapse, while the remaining channels may become highly correlated and encode redundant directions. Either mechanism reduces the usable feature space and weakens class separation. VICReg-style regularization [38] addresses both effects for global embeddings under clean augmentations. In SFDA-TAD, however, the constraints must operate on per-level foreground-background matrices because the tokens follow the detector's scale structure and the supervision remains noisy and incomplete.

Using the augmented level matrix $\tilde{Z}_l$ from Equation (4), we define the channel variance and decorrelation losses as follows.

$$\mathcal{L}_{\text{var}}^{(l)}(\tilde{Z}_l) = \frac{1}{C}\sum_{h=1}^{C} \max\left(0,\ \gamma - \sqrt{\text{Var}(\tilde{Z}_{l,:,h}) + \epsilon}\right), \tag{5}$$

$$\mathcal{L}_{\text{dec}}^{(l)}(\tilde{Z}_l) = \frac{1}{C(C-1)} \sum_{\substack{1 \le a,b \le C \\ a \ne b}} \text{Cov}(\tilde{Z}_l)_{ab}^2, \tag{6}$$

where $h \in \{1, \ldots, C\}$ indexes a feature channel, and $a, b \in \{1, \ldots, C\}$ index the entries of the channel covariance matrix. The vector $\tilde{Z}_{l,:,h}$ contains channel h across the $N_l + n_{\text{borrow}}$ tokens, while $\text{Var}(\cdot)$ and $\text{Cov}(\cdot)$ denote empirical variance and channel covariance, respectively. The threshold $\gamma > 0$ specifies the minimum channel standard deviation, and $\epsilon > 0$ provides numerical stability. Equation (5) penalizes a channel only when its standard deviation falls below γ, pushing low-variance channels away from constant responses while leaving sufficiently variable channels unchanged. Equation (6) minimizes squared off-diagonal covariance, encouraging different channels to encode non-redundant directions without constraining their individual variances. Together, the two losses preserve a non-degenerate feature space for the adaptive weighting mechanism described next.

3.6. Rank-Gap Reweighting

The channel losses above recover feature diversity, but uniform weighting cannot account for the level-dependent rank changes shown in Figure 1. Given the augmented matrix $\tilde{Z}_l$, RGR first computes the normalized effective rank r_l from its singular-value distribution. It then measures the gap between r_l and its smoothed running maximum over previous adaptation steps and converts this gap into the adaptive weight w_l. The resulting weight scales both channel losses at level l, assigning stronger regularization to levels with greater rank degradation. The effective rank and adaptive weight are defined in Equations (7) and (8), respectively.

$$r_l = \frac{100}{K_l} \exp\left(- \sum_{k=1}^{K_l} p_{l,k} \log p_{l,k} \right), \qquad p_{l,k} = \frac{\nu_{l,k}}{\sum_{u=1}^{K_l} \nu_{l,u}}, \tag{7}$$

$$w_l = 1 + \kappa \cdot \mathrm{clip}\left(\frac{\bar{r}_l^{\max} - r_l}{\bar{r}_l^{\max}}, \, 0, \, 1 \right), \tag{8}$$

where $K_l = \min(N_l + n_{\text{borrow}}, C)$ is the maximum possible rank of $\tilde{Z}_l$, $\nu_{l,k}$ is its k-th singular value, and $p_{l,k}$ is the corresponding normalized singular-value mass. The index u is used for normalization, with $0 \log 0 = 0$. The quantity $\bar{r}_l^{\max}$ is the exponentially smoothed running maximum of r_l over previous adaptation steps, $\kappa \geq 0$ controls the maximum increase in regularization strength, and $\mathrm{clip}(z, 0, 1) = \min(1, \max(0, z))$. The factor $100/K_l$ expresses effective rank as a percentage of its maximum possible value, and the resulting weight satisfies $1 \leq w_l \leq 1 + \kappa$.

Normalization by the maximum possible rank places matrices of different nominal sizes on a common scale. RGR uses this statistic to allocate regularization within each level, with the token population defined by DAS and CLB. Effective rank summarizes feature diversity rather than detection accuracy; the detection objective continues to determine action classification and temporal localization.

The historical reference is maintained separately for each level during target adaptation. It retains the level's previously attained feature diversity while preserving the distinct temporal roles of the pyramid levels. A reference that closely tracks the current rank can move downward with a sustained decline and attenuate the contrast used to allocate recovery strength. The smoothed maximum retains this contrast, and clipping bounds the resulting increase in channel regularization. The current weight uses the reference from previous steps, after which the reference is updated. No source recordings or target annotations are required.

Figure 1 illustrates why the allocation signal is separated from the task objective. At Level 6 in PGO-S2D, Vanilla Mean-Teacher has higher effective rank than both PLAID and the target-supervised oracle. PLAID therefore uses the within-level evolution of rank to distribute regularization while the detection loss optimizes the task; it does not seek the largest rank or a prescribed cross-level spectrum.

3.7. Overall Training Objective

As shown in Equation (9), we combine the variance and decorrelation losses across all pyramid levels under the adaptive weights w_l to form the PLAID regularization loss. The final training objective consists of the standard detection loss and the warmup-weighted PLAID loss, as shown in Equation (10).

$$\mathcal{L}_{\text{PLAID}} = \sum_{l=0}^{L-1} w_l \cdot \left(\alpha \cdot \mathcal{L}_{\text{var}}^{(l)}(\tilde{Z}_l) + \beta \cdot \mathcal{L}_{\text{dec}}^{(l)}(\tilde{Z}_l) \right). \tag{9}$$

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda(t) \cdot \mathcal{L}_{\text{PLAID}}, \qquad \lambda(t) = \lambda_{\max} \min\left(\frac{t}{T_{\text{w}}}, 1 \right), \tag{10}$$

where $\alpha \geq 0$ and $\beta \geq 0$ control the contributions of the variance and decorrelation losses, respectively, and $\lambda(t)$ controls the contribution of the PLAID loss to the final objective. Here, t is the training iteration, $\lambda_{\max} \geq 0$ is the terminal regularization coefficient, and $T_{\text{w}} > 0$ is the number of warmup updates. The warmup introduces the PLAID loss gradually, reducing the risk that strong representation regularization destabilizes early adaptation. The student is optimized using $\mathcal{L}$, whereas Vanilla MT uses only $\mathcal{L}_{\text{det}}$. The intermediate variant applies the variance and decorrelation losses with uniform weights and no token borrowing, replacing $\tilde{Z}_l$ in Equation (9) with Z_l. The full PLAID model includes all four components. Algorithm 1 summarizes one training epoch.

The additional computation in this formulation consists of token construction, empirical channel statistics, singular-value computation, and scalar reweighting during adaptation. For an augmented matrix with $M_l = N_l + n_{\text{borrow}}$ rows and C channels, forming a dense channel covariance costs on the order of $M_l C^2$ operations. Computing its singular-value spectrum from the feature matrix by a dense thin SVD has an order of $M_l C \min(M_l, C)$. Together, these operations provide level-specific regularization during source-free target adaptation.

DAS, CLB, and RGR reuse the existing pyramid features without introducing an additional trainable detector branch. Evaluation uses the EMA teacher with the detector architecture described in Section 3.1. The additional representation operations are used during adaptation. Measured timing and memory, together with the detector parameter and arithmetic counts, are reported in the experiments.

Algorithm 1 PLAID training for one epoch.

Input Student θ, teacher $\bar{\theta}$, target data $\mathcal{D}_t$, confidence threshold τ
for each batch $\{x_i\}$ from $\mathcal{D}_t$ **do**
  Run teacher forward pass to obtain proposals and scores
  Filter predictions by τ and apply Soft-NMS to form pseudo-segments
  Compute detection loss $\mathcal{L}_{\text{det}}$ on pseudo-segments
  Dispatch pseudo-segments to pyramid levels and sample tokens to obtain $\{Z_l\}$
  for each level $l = 0, \ldots, L-1$ **do**
    Borrow adjacent-level tokens and construct $\tilde{Z}_l$
    Compute channel losses and current effective rank r_l
    Compute w_l using the rank reference from previous steps
    Update the smoothed rank reference $\bar{r}_l^{\max}$
  end for
  Compute $\mathcal{L}_{\text{PLAID}}$ from all levels
  Update student using the total objective in Equation (10)
  Update teacher by exponential moving average
end for
Output Teacher weights $\bar{\theta}$ for evaluation

4. Experiments

The motivation analyses in Figures 1 and 2 establish level-dependent rank degradation and the limited ability of confidence filtering to provide pseudo-labels that are both reliable and complete. The experiments examine whether PLAID translates these observations into measurable adaptation gains under controlled duration and density shifts. Ablations assess the contribution of individual components and the proposed interaction between RGR and CLB. We further evaluate sensitivity to regularization strength, transfer across detector architectures, and generalization beyond the PGO benchmark on THUMOS14.

4.1. Dataset and Implementation

Experiments use the Power Grid Operations (PGO) dataset, collected for this study from fixed-camera recordings of real-world on-site operations. It contains 9800 videos with temporal annotations for five actions in the operational procedure: utility-pole ascent (up), utility-pole descent (down), verification, grounding-rod placement, and grounding-rod removal. Table 1 summarizes the dataset statistics.

The recordings cover more than 20 distinct monitoring scenes and involve more than 50 operators. Specific recording locations and operator identities are not disclosed to protect privacy and operational confidentiality.

Table 1. Statistics of the PGO dataset under four source-to-target transfer settings. $\bar{d}_{\text{act}}$ denotes the mean duration (s) of annotated action segments and defines the PGO-S2L/L2S partitions, which are based on action rather than video duration. $\bar{\rho}_{\text{act}}$ denotes the mean per-video action occupancy and defines the PGO-S2D/D2S partitions. T/S is the target-to-source ratio of the corresponding partition statistic. Videos and Segments report dataset sizes. Ground-truth (GT) action-segment distribution is the percentage of annotated segments belonging to each class. Verif., Put, and Take denote verification, putting a grounding rod, and taking a grounding rod, respectively.

Setting	Domain	Videos	Segments	Statistic	Value	T/S	GT Action-Segment Distribution (%)				
							Up	Verif.	Put	Down	Take
PGO-S2L	Source	4914	2513	$\bar{d}_{\text{act}}$ (s)	22.4	2.96	30.4	19.5	10.9	30.6	8.6
	Target	4886	3582	$\bar{d}_{\text{act}}$ (s)	66.3		31.9	10.9	16.9	28.5	11.8
PGO-L2S	Source	4954	3580	$\bar{d}_{\text{act}}$ (s)	66.3	0.34	31.6	11.2	17.0	28.4	11.8
	Target	4846	2515	$\bar{d}_{\text{act}}$ (s)	22.4		30.7	19.0	10.9	30.7	8.7
PGO-S2D	Source	4914	2347	$\bar{\rho}_{\text{act}}$	0.179	1.96	30.5	15.9	14.5	28.5	10.6
	Target	4886	3748	$\bar{\rho}_{\text{act}}$	0.350		31.8	13.5	14.4	29.9	10.4
PGO-D2S	Source	4978	3783	$\bar{\rho}_{\text{act}}$	0.347	0.52	30.6	13.9	14.7	30.6	10.3
	Target	4822	2312	$\bar{\rho}_{\text{act}}$	0.179		32.4	15.4	14.1	27.4	10.8

We construct four source-to-target settings along two orthogonal temporal shifts. For duration shift, $\bar{d}_{\text{act}}$ denotes the mean annotated action-segment duration, and $R_{\text{dur}} = \bar{d}^{\text{tgt}}_{\text{act}}/\bar{d}^{\text{src}}_{\text{act}}$ equals 2.96 for PGO-Short-to-Long (PGO-S2L) and 0.34 for PGO-Long-to-Short (PGO-L2S); these settings are partitioned by action rather than video duration. For density shift, $\bar{\rho}_{\text{act}}$ denotes the mean per-video action occupancy, and $R_{\text{den}} = \bar{\rho}^{\text{tgt}}_{\text{act}}/\bar{\rho}^{\text{src}}_{\text{act}}$ equals 1.96 for PGO-Sparse-to-Dense (PGO-S2D) and 0.52 for PGO-Dense-to-Sparse (PGO-D2S). Each name indicates the source-to-target direction, and ratios farther from one indicate stronger shifts. Duration and density shifts affect the regression and classification priors, respectively.

Source and target domains are constructed from non-overlapping recording-site groups. Videos are selected within their respective site groups according to action duration or action occupancy. All clips and annotated action segments derived from the same original recording retain that recording's domain assignment. The temporal statistics characterize the domain shift, while recording and site identities enforce partition independence.

Unless otherwise stated, experiments use ActionFormer [1], a six-level temporal-pyramid detector with a PointGenerator head, and pre-extracted 2048-dimensional RGB-stream I3D features [26] pretrained on Kinetics-400 with snippet stride 4. Each source model is trained for 80 epochs on two GPUs using AdamW, a learning rate of 10^{-4} with cosine decay, and batch size 2; the resulting checkpoint initializes target adaptation. Single-run baseline results use random seed 42. Mean-Teacher denotes the source-free adaptation framework, whereas ActionFormer and TemporalMaxer denote the underlying TAD architectures; the cross-architecture evaluation substitutes TemporalMaxer for ActionFormer under the same adaptation setting.

By default, target adaptation runs for 10 epochs using AdamW with batch size 2, a learning rate of 10^{-5}, confidence threshold $\tau = 0.5$, and teacher momentum $\mu = 0.999$. Gaussian Soft-NMS uses bandwidth $\sigma_{\text{NMS}} = 0.5$ and an IoU threshold of 0.1. PLAID uses $\alpha = 0.5$, $\beta = 0.5$, $\gamma = 1.0$, $\kappa = 1.0$, $n_{\text{fg}} = 8$, $n_{\text{bg}} = 32$, $n_{\text{borrow}} = 16$, terminal coefficient $\lambda_{\max} = 10$, and warmup length $T_{\text{w}} = 50$ updates. We evaluate sensitivity to $\lambda_{\max}$ in Figure 4. Performance is reported as mean average precision (mAP) at temporal intersection-over-union (tIoU) thresholds from 0.1 to 0.6 and their arithmetic mean. The final EMA teacher is evaluated on a predefined 1000-video target subset, with no target labels used for checkpoint selection.

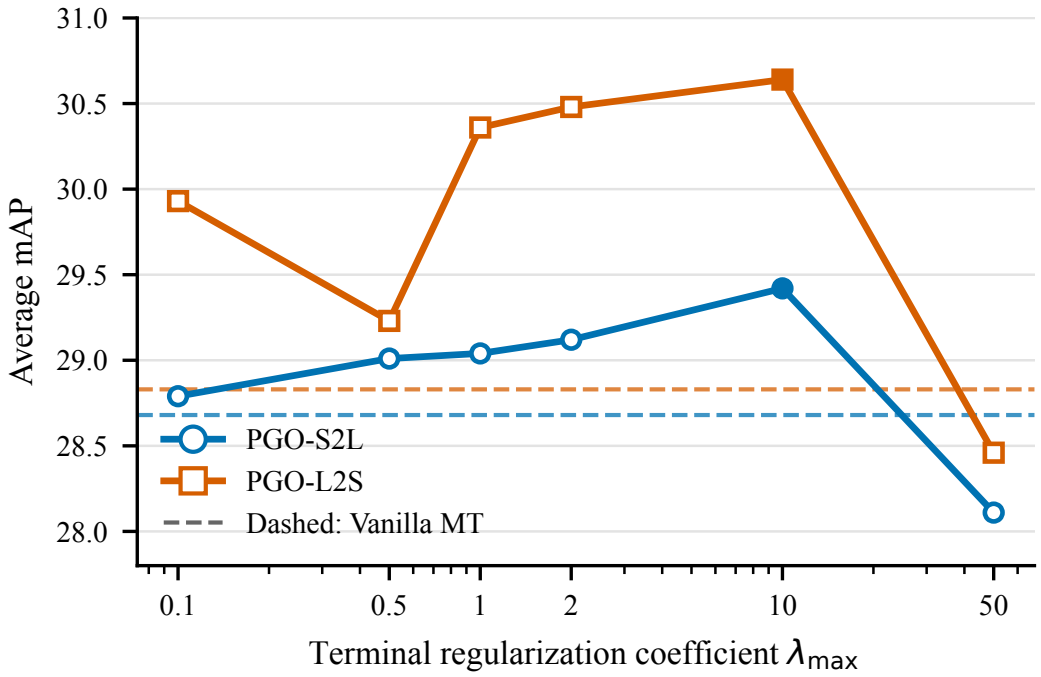


Figure 4. Sensitivity to the terminal regularization coefficient λ_{max} with ActionFormer in PGO-S2L and PGO-L2S. The dashed lines denote the corresponding Vanilla Mean-Teacher references. Both settings peak at $\lambda_{max} = 10$, whereas $\lambda_{max} = 50$ falls below the reference in each setting.

Vanilla Mean-Teacher (Vanilla MT) denotes the standard Mean-Teacher adaptation baseline, whereas No-EMA denotes the same setting with the EMA teacher update disabled. The transferred baselines are Feature Augmentation Based Test-Time Adaptation (FATA) [43], Source Hypothesis Transfer with information maximization (SHOT-IM) [4], Neighborhood Reciprocity Clustering (NRC) [44], Attracting and Dispersing (AaD) [45], and Dual-Rate Dynamic Teacher (DDT) [46].

4.2. Main Results

Tables 2 and 3 report mAP at each IoU threshold for the absolute-action-duration and action-density dimensions. Adaptation is not uniformly beneficial in this regime. Vanilla Mean-Teacher is the only off-the-shelf method that exceeds source-only in all four PGO settings, with average gains ranging from 1.31 mAP in PGO-S2L to 4.03 mAP in PGO-L2S. PLAID further improves over Vanilla Mean-Teacher in every setting. Averaged over the six thresholds, the gains over source-only are 2.50 mAP in PGO-S2L, 6.03 in PGO-L2S, 2.87 in PGO-S2D, and 3.98 in PGO-D2S. The gains over Vanilla Mean-Teacher are 1.19 mAP in PGO-S2L, 2.00 in PGO-L2S, 1.37 in PGO-S2D, and 0.45 in PGO-D2S.

DDT exhibits a pronounced directional asymmetry. It improves over source-only by 0.44 mAP in PGO-L2S and 1.50 mAP in PGO-D2S, but falls below source-only by 11.85 mAP in PGO-S2L and 13.51 mAP in PGO-S2D. It also remains below Vanilla Mean-Teacher in all four settings. These results show that DDT does not provide consistent gains across the duration and density shifts considered here, with larger degradation when the target domain contains longer or denser actions.

The target-lighter settings, PGO-L2S and PGO-D2S, show larger absolute gains over source-only, whereas the target-heavier settings, PGO-S2L and PGO-S2D, show more limited improvement. This directional asymmetry suggests that adaptation becomes more difficult when the target distribution extends toward longer-duration or denser actions. The per-IoU results qualify the average gains. In PGO-L2S, PLAID exceeds Vanilla Mean-Teacher at every threshold by 0.53 to 2.84 mAP, with the largest margin at tIoU 0.4. In PGO-D2S,

the average margin is smaller; PLAID is lower at tIoU 0.1 and 0.3 and higher at the other evaluated thresholds.

Table 2. ActionFormer-based results for PGO-S2L and PGO-L2S. The columns from 0.1 to 0.6 report mAP at the corresponding temporal intersection-over-union (tIoU) threshold using all samples in the predefined target evaluation subset without weighting by class frequency, and Avg. is their arithmetic mean. FATA, No-EMA, and SHOT-IM perform below source-only in both settings; NRC and AaD remain close to source-only in PGO-L2S but fall below it in PGO-S2L. For Vanilla MT and PLAID, the three adaptation seeds are 42, 43, and 44. Threshold-specific entries and Avg. are three-seed means; the corresponding standard deviations are reported in Appendix A Table A1. Other methods are reported as single runs. Bold type identifies PLAID.

PGO Setting	Method	mAP at tIoU						Avg.
		0.1	0.2	0.3	0.4	0.5	0.6	
PGO-S2L	Source-only	32.53	31.88	30.00	27.17	23.28	17.36	27.04
PGO-S2L	No-EMA	24.04	18.83	15.87	12.94	11.52	10.79	15.66
PGO-S2L	FATA [43]	22.42	20.14	18.35	15.49	11.79	8.41	16.10
PGO-S2L	SHOT-IM [4]	16.23	14.13	13.07	11.05	8.03	7.02	11.59
PGO-S2L	NRC [44]	27.94	24.62	22.95	19.74	18.82	16.65	21.79
PGO-S2L	AaD [45]	28.30	23.14	22.10	20.19	18.07	17.03	21.47
PGO-S2L	DDT [46]	21.32	19.64	17.34	14.47	11.09	7.30	15.19
PGO-S2L	Vanilla MT [5]	35.14	35.11	31.99	27.45	23.22	17.21	28.35
PGO-S2L	**PLAID**	37.20	35.23	32.89	26.88	25.22	19.80	29.54
PGO-L2S	Source-only	29.60	29.10	27.46	24.55	22.04	16.80	24.93
PGO-L2S	No-EMA	23.67	21.65	19.56	17.30	14.02	11.61	17.97
PGO-L2S	FATA [43]	30.76	30.14	28.72	26.22	19.43	12.77	24.67
PGO-L2S	SHOT-IM [4]	14.11	13.26	13.08	11.04	10.02	9.02	11.75
PGO-L2S	NRC [44]	28.63	28.12	27.41	26.18	21.83	18.15	25.05
PGO-L2S	AaD [45]	28.16	27.07	26.05	24.04	20.03	16.01	23.56
PGO-L2S	DDT [46]	31.76	30.94	29.52	25.12	20.93	13.92	25.37
PGO-L2S	Vanilla MT [5]	35.80	34.55	31.66	28.71	24.22	18.80	28.96
PGO-L2S	**PLAID**	37.90	36.71	34.01	31.55	26.23	19.33	30.96

Table 3. ActionFormer-based results for PGO-S2D and PGO-D2S. The columns from 0.1 to 0.6 report mAP at the corresponding temporal intersection-over-union (tIoU) threshold using all samples in the predefined target evaluation subset without weighting by class frequency, and Avg. is their arithmetic mean. Most image-classification SFDA baselines fall below source-only; FATA exceeds it only in PGO-D2S. For Vanilla MT and PLAID, the three adaptation seeds are 42, 43, and 44. Threshold-specific entries and Avg. are three-seed means; the corresponding standard deviations are reported in Appendix A Table A1. Other methods are reported as single runs. Bold type identifies PLAID.

PGO Setting	Method	mAP at tIoU						Avg.
		0.1	0.2	0.3	0.4	0.5	0.6	
PGO-S2D	Source-only	38.10	37.00	35.89	33.10	29.17	21.63	32.48
PGO-S2D	No-EMA	22.98	21.66	19.83	16.89	13.55	11.18	17.68
PGO-S2D	FATA [43]	28.73	26.29	23.17	20.22	16.11	10.89	20.90
PGO-S2D	SHOT-IM [4]	26.09	23.04	21.03	20.01	19.02	17.02	21.04
PGO-S2D	NRC [44]	27.65	26.38	23.28	22.25	19.17	17.11	22.64
PGO-S2D	AaD [45]	23.18	21.78	20.12	17.52	14.66	12.13	18.23
PGO-S2D	DDT [46]	25.23	24.27	21.64	19.34	14.61	8.73	18.97
PGO-S2D	Vanilla MT [5]	41.50	39.75	37.33	33.83	29.55	21.93	33.98
PGO-S2D	**PLAID**	42.98	41.66	39.83	35.89	30.55	21.18	35.35

Table 3. *Cont.*

PGO Setting	Method	mAP at tIoU						Avg.
		0.1	0.2	0.3	0.4	0.5	0.6	
PGO-D2S	Source-only	27.02	26.34	24.99	23.05	20.50	15.32	22.87
PGO-D2S	No-EMA	21.21	18.81	16.04	13.55	11.44	10.35	15.23
PGO-D2S	FATA [43]	31.21	29.44	26.72	23.21	20.18	12.93	23.95
PGO-D2S	SHOT-IM [4]	22.06	21.04	20.98	18.02	17.01	12.01	18.52
PGO-D2S	NRC [44]	21.63	19.32	16.66	14.69	12.01	10.63	15.82
PGO-D2S	AaD [45]	20.70	18.91	18.19	17.11	16.05	15.01	17.66
PGO-D2S	DDT [46]	30.80	29.57	27.82	24.51	19.78	13.73	24.37
PGO-D2S	Vanilla MT [5]	32.30	31.30	29.50	26.50	22.78	16.00	26.40
PGO-D2S	**PLAID**	32.22	31.33	29.10	26.97	24.22	17.24	26.85

The comparison with image-classification source-free baselines connects the main results to the pseudo-label motivation in Figure 2. No-EMA and SHOT-IM fall below source-only in every setting, with SHOT-IM reaching 14.11 mAP at tIoU 0.1 in PGO-L2S compared with 29.60 for source-only. NRC and AaD also fall below source-only in PGO-S2L, PGO-S2D, and PGO-D2S. FATA underperforms source-only in PGO-S2L, PGO-L2S, and PGO-S2D, but exceeds it by 1.08 mAP in PGO-D2S. In PGO-L2S, both NRC and FATA remain within 0.26 mAP of source-only. These results are consistent with the precision–recall analysis, in which removing noise through confidence filtering also discards a substantial fraction of usable predictions. The contrast between No-EMA and Vanilla Mean-Teacher is consistent with stabilization from exponential moving averaging, while PLAID additionally regularizes the representation after pseudo-label selection. Among the tested implementations, only the Mean-Teacher-based variants exceed source-only average mAP in all four PGO settings.

Fine-Grained and Statistical Evaluation

The paired-seed comparison in Table A1 uses three adaptation seeds, 42, 43, and 44, and separates training variability from threshold-wise averaging. Each run is first averaged over the six tIoU thresholds; method summaries use the across-seed mean and sample standard deviation, and the comparison uses differences between matched seeds. Cross-direction pseudo-label diagnostics use the deployed confidence threshold, Soft-NMS, and class-aware one-to-one temporal matching, as specified in Table 4.

Table 4. Cross-direction pseudo-label diagnostics from the source-only checkpoint before target adaptation, evaluated on the predefined target evaluation subset at the deployed confidence threshold and a tIoU matching threshold of 0.5. Predictions are evaluated after confidence filtering and Soft-NMS with class-aware, one-to-one matching to target annotations. Target labels are used only for this offline diagnostic. Precision, recall, and F1 are percentages.

Setting	Confidence	tIoU	Precision	Recall	F1
PGO-S2L	0.5	0.5	12.2	13.3	12.7
PGO-L2S	0.5	0.5	11.7	14.2	12.8
PGO-S2D	0.5	0.5	29.9	18.4	22.8
PGO-D2S	0.5	0.5	18.1	17.2	17.6

Class-wise AP and duration-stratified mAP examine whether gains are concentrated in particular actions or temporal scales, as organized in Tables A2 and A3. Duration is measured for each annotated action segment in seconds, rather than for the complete video. The shared groups are at most 10 s, more than 10 and at most 30 s, and more than 30 s. Out-of-bin annotations are ignored during bin-specific matching.

In the single-run class-wise results in Table A2, PLAID exceeds Vanilla MT in all five action categories for PGO-S2L and PGO-S2D. For PGO-S2L, the gains range from 0.39 AP for grounding-rod removal to 1.35 AP for grounding-rod placement. For PGO-S2D, the gains range from 1.45 AP for grounding-rod placement to 1.59 AP for grounding-rod removal. In PGO-D2S, PLAID improves the Up, Down, Put and Take categories, while the verification AP is lower than that of Vanilla MT.

Table A3 reports results by action duration. PLAID exceeds Vanilla MT in each of the three duration groups in all four transfer settings. The observed gains range from 0.14 to 2.24 mAP across the reported groups, with the largest gain occurring for actions of 10–30 s in PGO-L2S. The groups contain different numbers of represented classes, as reported in the table.

Table 5 reports a matched fixed-input resource profile that separates adaptation updates from detector inference. The paired-seed gain statistics, class-wise and duration-stratified results, and additional parameter comparisons are reported in Appendix A, Tables A1–A4.

Table 5. Fixed-input resource profile. Adapt. is mean adaptation time per update, rather than full-run wall time. The last two columns report detector-forward latency and forward plus Soft-NMS latency per feature window. GMACs count convolution and dense matrix products in one detector forward pass; normalization, elementwise operations, NMS, and I3D are excluded from this arithmetic count. The counted operations require 361.2 GFLOPs (1 MAC = 2 FLOPs). PLAID adds no trainable parameters or inference branch.

Method	Adapt. (ms)	Peak (GiB)	Params (M)	GMACs	Forward (ms)	+NMS (ms)
Vanilla MT	383.76	4.841	29.23	180.59	56.30	57.34
PLAID	424.50	4.962	29.23	180.59	57.10	57.48

Runtime measurements use an NVIDIA A800-SXM4-80GB, PyTorch 2.3.0 with CUDA 12.1, and FP32 with TF32 disabled. Input feature windows contain 2304 tokens with 2048 channels; the detector configuration pads them internally to 9216 tokens. Adaptation uses batch size 2 and inference uses batch size 1. Adaptation timing includes teacher inference, pseudo-label filtering and Soft-NMS, student forward and backward passes, AdamW, and EMA updates. It averages 12 measured updates after three warmup updates on a fixed real-feature batch. Inference averages 30 runs after five warmup runs. Data loading and I3D extraction are excluded. The PLAID profiling implementation follows the manuscript objective with a total borrowing budget of 16 and the stated regularization coefficients. Peak memory is the PyTorch peak allocated memory on the measured device.

4.3. Ablation Study

Table 6 isolates each PLAID component in PGO-L2S. Source-only obtains 24.93 mAP, Vanilla Mean-Teacher obtains 28.83, and the full PLAID model obtains 30.64. The variant containing DAS, variance, and decorrelation but neither RGR nor CLB reaches 29.56 mAP, which is 0.73 mAP above Vanilla Mean-Teacher. Removing DAS, variance, or decorrelation from the full model reduces performance to 30.11, 30.31, and 30.09 mAP, respectively. These comparable reductions show that each foundation component contributes to the full model, with the largest observed decrease caused by removing decorrelation.

The component interaction should be interpreted relative to the appropriate reference. Starting from the foundation variant at 29.56 mAP, adding RGR without CLB reduces performance to 28.89, a decrease of 0.67 mAP. This remains 0.06 mAP above Vanilla Mean-Teacher at 28.83. Adding CLB without RGR gives 29.58, whereas their joint use gives 30.64, which is 1.08 mAP above the foundation variant. The observed gain is therefore associated with the joint configuration rather than an independent benefit from either component.

RGR allocates the strength of both channel penalties according to the estimated rank gap. Limited token support can make that allocation sensitive to sampling fluctuations. CLB supplies temporally corresponding adjacent-level tokens before the rank and channel statistics are computed, supporting the statistics on which RGR operates. This provides a mechanism-level explanation consistent with the observed ablation pattern.

The joint configuration addresses both statistical support and regularization allocation. CLB changes the population used to estimate the statistics, while RGR determines how strongly the resulting channel constraints act at each level. Table 6 supports this coupled design in PGO-L2S.

Table 6. Ablation in PGO-L2S (seed 42). Individual component removals and the RGR and CLB interaction are shown. Δ is the change in average mAP against the source-only baseline of 24.93. Checkmarks indicate included components and dashes indicate excluded components or a non-applicable comparison. Bold indicates the best result in each numerical column.

Variant	DAS	$\mathcal{L}_{\mathbf{var}}$	$\mathcal{L}_{\mathbf{dec}}$	RGR	CLB	mAP	Δ vs. Source-Only
Source-only	–	–	–	–	–	24.93	–
Vanilla MT	–	–	–	–	–	28.83	3.90
w/o DAS	–	✓	✓	✓	✓	30.11	5.18
w/o $\mathcal{L}_{\text{var}}$	✓	–	✓	✓	✓	30.31	5.38
w/o $\mathcal{L}_{\text{dec}}$	✓	✓	–	✓	✓	30.09	5.16
w/o RGR & CLB	✓	✓	✓	–	–	29.56	4.63
w/o CLB	✓	✓	✓	✓	–	28.89	3.96
w/o RGR	✓	✓	✓	–	✓	29.58	4.65
PLAID full	✓	✓	✓	✓	✓	**30.64**	**5.71**

4.4. Sensitivity to Regularization Strength

We evaluate the sensitivity of PLAID to the terminal regularization coefficient $\lambda_{\max}$ in both duration-shift settings. As shown in Figure 4, both settings peak at $\lambda_{\max} = 10$, reaching 29.42 mAP in PGO-S2L and 30.64 in PGO-L2S. Every setting from $\lambda_{\max} = 0.1$ to $\lambda_{\max} = 10$ exceeds the corresponding Vanilla Mean-Teacher baseline, which is 28.68 mAP in PGO-S2L and 28.83 in PGO-L2S. At $\lambda_{\max} = 50$, performance decreases to 28.11 in PGO-S2L and 28.46 in PGO-L2S, below the corresponding baseline in both settings. The shared reduction is consistent with excessive regularization overwhelming the detection objective.

The response to $\lambda_{\max}$ differs by shift direction. PGO-L2S spans 1.41 mAP between $\lambda_{\max} = 0.5$ and $\lambda_{\max} = 10$, whereas PGO-S2L spans only 0.41 mAP over the same interval. PGO-L2S is therefore more responsive to the regularization weight, while PGO-S2L has a flatter response. At $\lambda_{\max} = 0.1$, both settings remain above their respective Vanilla Mean-Teacher baselines, although the margin in PGO-S2L is only 0.11 mAP. Overall, the interval from $\lambda_{\max} = 0.5$ to $\lambda_{\max} = 10$ is stable in both tested settings, while $\lambda_{\max} = 50$ represents a shared over-regularization regime.

The other parameters control different aspects of the objective. The variance and decorrelation coefficients set the relative contribution of the two channel penalties, while the standard-deviation floor determines which low-variance channels are penalized. The rank-gap coefficient bounds the additional emphasis assigned to a level. Foreground, background, and borrowing budgets determine the composition and nominal size of the statistical sample, and the warmup length controls how quickly the representation penalty reaches its full strength. The present sweep isolates the terminal regularization strength with these quantities fixed. Appendix A Table A4 separates the effects of rank-gap weighting and the cross-level borrowing budget through one-factor comparisons.

Table A4 compares the additional parameters in PGO-L2S and PGO-S2D. For the tested rank-gap coefficients, average mAP increases from 29.58 and 33.21 at $\kappa = 0$ to 30.64 and 35.23 at $\kappa = 1$, respectively, and reaches 30.44 and 35.02 at $\kappa = 2$. For the borrowing budget, the default of 16 gives the highest result among the tested values in both settings. At a budget of 32, average mAP is 29.88 in PGO-L2S and 35.20 in PGO-S2D.

4.5. Cross-Architecture Effectiveness

Table 7 evaluates whether PLAID remains effective when ActionFormer is replaced by TemporalMaxer in PGO-L2S and PGO-D2S. PLAID exceeds both source-only and Vanilla Mean-Teacher at every evaluated threshold in both settings. In PGO-L2S, its average mAP is 21.65, compared with 19.79 for source-only and 21.35 for Vanilla Mean-Teacher, yielding gains of 1.86 and 0.30 mAP, respectively. In PGO-D2S, PLAID reaches 18.57 mAP, compared with 17.94 and 18.22, yielding gains of 0.63 and 0.35 mAP. These consistent improvements demonstrate that PLAID retains its effectiveness across ActionFormer and TemporalMaxer.

Table 7. TemporalMaxer-based results for the target-lighter PGO-L2S and PGO-D2S settings. The columns from 0.1 to 0.6 report mAP at the corresponding tIoU threshold, and Avg. is their arithmetic mean. Here, TemporalMaxer replaces ActionFormer as the TAD detector, while Vanilla MT and PLAID remain adaptation methods applied to that detector. PLAID surpasses both source-only and Vanilla MT at every threshold, while No-EMA performs below source-only in both settings. Bold indicates the best result in each numerical column.

PGO Setting	Method	mAP at tIoU						Avg.
		0.1	0.2	0.3	0.4	0.5	0.6	
PGO-L2S	Source-only	24.31	23.69	22.15	19.47	16.73	12.39	19.79
	Vanilla MT	26.53	25.87	24.12	21.04	17.33	13.19	21.35
	No-EMA	17.50	14.41	10.32	8.26	6.21	5.04	10.29
	PLAID	**26.55**	**26.08**	**24.44**	**21.22**	**17.74**	**13.89**	**21.65**
PGO-D2S	Source-only	23.11	22.06	20.27	18.12	14.19	9.89	17.94
	Vanilla MT	23.51	22.52	20.78	18.03	14.78	9.70	18.22
	No-EMA	16.54	13.01	11.58	6.99	5.01	3.95	9.51
	PLAID	**23.85**	**22.94**	**21.09**	**18.46**	**14.99**	**10.07**	**18.57**

The TemporalMaxer result also supports the structural motivation of PLAID because both detectors organize temporal features through a multi-level pyramid. Self-training without EMA stabilization remains ineffective, with No-EMA decreasing to approximately 10.3 mAP in PGO-L2S and 9.5 in PGO-D2S. The positive margins of PLAID are smaller than those obtained with ActionFormer, including a reduction from 2.00 to 0.30 mAP over Vanilla Mean-Teacher in PGO-L2S. TemporalMaxer also has lower source-only averages of 19.79 mAP in PGO-L2S and 17.94 in PGO-D2S, compared with 24.93 and 22.87 for ActionFormer. This comparison suggests that the initial target-domain signal affects the magnitude of the adaptation gain, although it does not isolate the cause.

The threshold-resolved results strengthen the cross-architecture conclusion. PLAID retains a positive margin over both baselines from tIoU 0.1 through 0.6 in PGO-L2S and PGO-D2S, including the stricter thresholds of 0.5 and 0.6. The improvement therefore persists across a range of localization tolerances rather than arising from a favorable operating point. Together with the average mAP gains, this consistency provides evidence that the effectiveness of PLAID extends beyond the primary ActionFormer detector.

4.6. Generalization on THUMOS14

We further evaluate PLAID on THUMOS14 [47] to examine whether its effectiveness extends beyond the domain-specific PGO benchmark. We construct a Short-to-Long source-to-target transfer by partitioning videos according to the absolute durations of their annotated action segments rather than video duration. The resulting target domain has a mean action-segment duration 3.31 times that of the source domain, i.e., $R_{\mathrm{dur}} = \bar{d}^{\mathrm{tgt}}_{\mathrm{act}} / \bar{d}^{\mathrm{src}}_{\mathrm{act}} = 3.31$. All methods use ActionFormer with the same data split, feature representation, adaptation setting, and evaluation protocol. For consistency with the PGO evaluation, Table 8 reports mAP at tIoU thresholds from 0.1 to 0.6 and their arithmetic mean.

Table 8. ActionFormer-based results for the THUMOS14 Short-to-Long transfer constructed according to the absolute durations of annotated action segments. The target-domain mean action duration is 3.31 times that of the source domain. The columns from 0.1 to 0.6 report mAP at the corresponding tIoU threshold, and Avg. is their arithmetic mean. Bold indicates the best result in each numerical column.

Method	**mAP at tIoU**						**Avg.**
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	
Source-only	82.21	80.06	77.12	72.34	64.59	53.37	71.62
No-EMA	66.23	58.97	50.10	40.79	36.00	24.80	46.15
FATA [43]	83.29	79.24	72.67	65.46	53.70	41.11	65.91
SHOT-IM [4]	76.87	74.83	63.96	57.16	46.09	40.10	59.84
NRC [44]	77.81	74.94	70.29	61.84	51.82	38.13	62.47
AaD [45]	75.32	73.31	68.44	52.77	46.05	33.53	58.24
DDT [46]	80.70	78.55	73.71	67.25	56.94	42.78	66.66
Vanilla MT [5]	84.59	82.34	79.67	75.96	65.90	50.81	73.21
PLAID	**88.28**	**86.63**	**83.51**	**79.28**	**70.97**	**57.42**	**77.68**

PLAID achieves the best result at every evaluated threshold and reaches 77.68 average mAP. This exceeds source-only by 6.06 mAP and Vanilla Mean-Teacher by 4.47 mAP, despite the source-only detector already attaining 71.62 mAP. Most transferred adaptation baselines fall below source-only, whereas Vanilla Mean-Teacher yields a smaller improvement to 73.21 mAP. The advantage of PLAID over Vanilla Mean-Teacher increases to 5.07 and 6.61 mAP at tIoU 0.5 and 0.6, respectively, showing that the gain persists under stricter localization criteria. This result extends the evidence beyond PGO while retaining the same controlled absolute-duration shift.

4.7. Qualitative Case Study

Figure 5 provides a qualitative comparison of source-only and PLAID localization for utility-pole descent (down) and ascent (up). In the descent example, the source-only prediction separates the action into an early fragment and a later fragment, whereas PLAID recovers a single interval that follows the annotated action. The ascent example exhibits the same qualitative pattern: source-only leaves a substantial gap within the ground-truth interval, while PLAID produces a temporally coherent detection with closer start and end boundaries. These examples illustrate that the quantitative gains can correspond to reduced temporal fragmentation and improved boundary alignment in safety-critical operational actions. The two representative videos provide a qualitative view of the detection outputs.

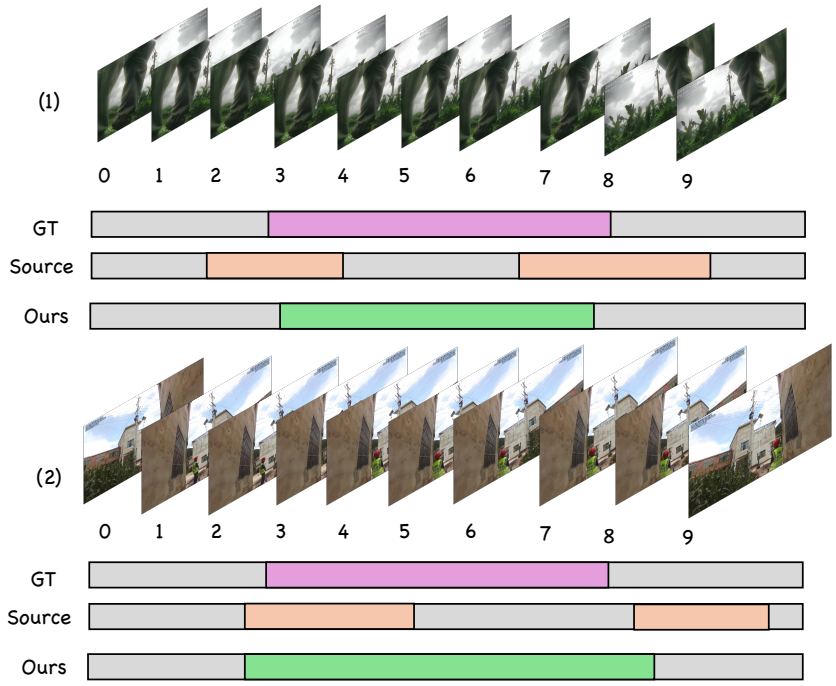


Figure 5. Qualitative localization of utility-pole descent (down) and ascent (up) in two target-domain videos. (1) shows descent and (2) shows ascent. Ground truth, source-only predictions, and PLAID predictions are shown in purple, orange, and green, respectively. PLAID reduces temporal fragmentation and produces more continuous intervals with boundaries closer to the ground truth.

5. Discussion

The central finding is that source-free adaptation is governed not by temporal-shift magnitude alone, but by its interaction with the quality of the initial target-domain signal. Within PGO, PLAID yields larger gains when transferring toward shorter or sparser actions than toward longer or denser actions. The THUMOS14 experiment argues against interpreting this asymmetry as an intrinsic limitation of longer-target transfer. Despite a 3.31-fold increase in mean action duration, PLAID improves over source-only by 6.06 mAP and over Vanilla Mean-Teacher by 4.47 mAP. This contrast suggests that temporal structure determines where the mismatch appears in the detector, whereas initial target performance constrains how much reliable signal remains available for adaptation. Cross-direction pseudo-label diagnostics provide additional context for these differences.

In fixed-camera power-grid monitoring, differences in operator pace, procedure duration, and inactive intervals can alter temporal statistics across deployment sites even when the camera configuration remains unchanged. The source-free setting is particularly relevant because operational recordings may expose sensitive infrastructure and cannot always be retained or transferred between regional utilities. PLAID addresses this deployment condition by adapting to site-specific temporal patterns without requiring access to source videos.

The results also clarify why image-oriented SFDA methods transfer unreliably to TAD. Temporal detection couples class prediction with boundary regression, so errors in either component can corrupt pseudo-label content and its assignment across pyramid levels. The level-dependent rank changes in Figure 1 provide representation-level evidence that this difficulty is not captured by confidence alone. PLAID therefore uses effective rank as a label-free signal for allocating regularization rather than as a quantity to maximize directly. In PGO-S2D, its movement toward the source-domain reference at the first five levels supports the intended preservation of feature diversity, while the reversed ordering at Level 6 and the target-supervised oracle show that no single rank profile is universally

optimal. The positive results with TemporalMaxer further demonstrate that the approach extends across both evaluated temporal-pyramid detector architectures and representative transfer settings.

The evaluation covers four duration and density shifts on PGO, one duration shift on THUMOS14, and two PGO transfers with TemporalMaxer. Effective rank provides a diagnostic of feature diversity alongside the detection results and component ablations. Together, these analyses support pyramid-level representation regularization as a complementary signal for source-free adaptation with imperfect pseudo-labels. Future work can examine additional datasets, transfer directions, and detector families.

6. Conclusions

PLAID addresses source-free domain adaptation in temporal action detection by complementing Mean-Teacher self-training with effective-rank-guided, pyramid-level representation regularization. Channel variance and decorrelation preserve feature diversity, while RGR and CLB allocate and stabilize recovery at degraded pyramid levels. Across four controlled duration and density shifts of PGO, PLAID improves source-only ActionFormer by 2.5–6.0 mAP and consistently outperforms Vanilla Mean-Teacher; the gains also persist with TemporalMaxer in two representative settings. On the THUMOS14 Short-to-Long transfer, PLAID reaches 77.68 average mAP, exceeding source-only and Vanilla Mean-Teacher by 6.06 and 4.47 mAP, respectively. The effective-rank analysis and ablations support the intended representation-recovery mechanism. For fixed-camera power-grid monitoring, PLAID provides a practical adaptation strategy when operational tempo and action occupancy vary across sites and sensitive source recordings cannot be shared. Broader transfer directions and detector families remain important directions for future evaluation.

Author Contributions: Conceptualization, L.M. and X.Z.; methodology, L.M. and X.Z.; software, L.M., S.Y., and M.X.; validation, L.M., S.Y., M.X., and X.Z.; formal analysis, L.M., S.Y., M.X., and X.Z.; investigation, L.M., S.Y., and M.X.; resources, S.Y. and M.X.; data curation, S.Y. and M.X.; writing—original draft preparation, L.M. and S.Y.; writing—review and editing, L.M., S.Y., M.X., and X.Z.; visualization, L.M., S.Y., and M.X.; supervision, L.M. and X.Z.; project administration, L.M. and X.Z.; funding acquisition, L.M. and X.Z. L.M. conceived the study and led the development of the PLAID framework, including Duration-Aware Sampling, Cross-Level Borrowing, channel variance and decorrelation regularization, and Rank-Gap Reweighting. L.M. coordinated the PGO duration and density shift experiments, the THUMOS14 evaluation, and the comparisons between ActionFormer and TemporalMaxer, and led the preparation of the original manuscript. S.Y. implemented the Duration-Aware Sampling and Cross-Level Borrowing components, prepared the duration and density transfer splits, and contributed to training, evaluation, ablation analyses, visualizations, and the original manuscript. M.X. implemented the Rank-Gap Reweighting and representation-diagnostic routines, curated and checked the PGO temporal annotations, and contributed to sensitivity analyses, comparisons across detectors, and the preparation of figures and result tables. S.Y. and M.X. supported data preparation and experimental resources. X.Z. co-developed the study design and methodology, contributed to the formal interpretation and validation of the representation-recovery mechanism, supervised the research, and coordinated project administration and funding acquisition with L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Guizhou Power Grid Co., Ltd., grant number GZKJXM20222320.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The public THUMOS14 benchmark data used in this study are available from the cited sources. The industrial Power Grid Operations (PGO) data are not publicly available due to privacy and operational restrictions.

Acknowledgments: During manuscript preparation and revision, Codex (OpenAI) was used to assist with manuscript organization, language revision, bibliographic checks, methodological clarifications, drafting responses to reviewers, and preparation of the Author Contributions and Conflicts of Interest statements. The authors remain responsible for verifying all statements, references, and results.

Conflicts of Interest: Authors Lingwen Meng, Siwu Yu, and Mingyong Xin were employed by the Electric Power Research Institute of Guizhou Power Grid Co., Ltd. The remaining author, Xinshan Zhu, declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The authors declare that this study received funding from Guizhou Power Grid Co., Ltd., grant number GZKJXM20222320. The funder was involved in the study through its employee authors, whose contributions are described in the Author Contributions statement.

Appendix A. Additional Evaluation

Table A1. Paired-seed comparison of Vanilla Mean-Teacher and PLAID using the three adaptation seeds 42, 43, and 44. Each run is first averaged over tIoU thresholds 0.1 through 0.6. Method entries report the across-seed mean and sample standard deviation. Δ is the mean of the within-seed differences, PLAID minus Vanilla MT. n is the number of matched seeds.

Setting	Vanilla MT	PLAID	Δ	n
PGO-S2L	28.35 ± 0.34	29.54 ± 0.16	1.19	3
PGO-L2S	28.96 ± 0.10	30.96 ± 0.35	2.00	3
PGO-S2D	33.98 ± 0.24	35.35 ± 0.15	1.37	3
PGO-D2S	26.40 ± 0.16	26.85 ± 0.14	0.45	3

Table A2. Class-wise AP averaged over tIoU thresholds 0.1 through 0.6. Entries report single-run results from the corresponding evaluated checkpoints. Verif., Put, and Take denote verification, grounding-rod placement, and grounding-rod removal. Overall average mAP is reported in Tables 2 and 3; those tables use three-seed averages for Vanilla MT and PLAID.

Setting	Method	Up	Down	Verif.	Put	Take
PGO-S2L	Source-only	25.44	53.59	11.85	22.76	13.96
PGO-S2L	Vanilla MT	37.03	55.12	13.34	23.77	15.89
PGO-S2L	PLAID	37.86	56.08	14.22	25.12	16.28
PGO-L2S	Source-only	33.18	51.33	9.77	20.21	10.11
PGO-L2S	Vanilla MT	36.92	55.76	11.29	24.52	14.76
PGO-L2S	PLAID	38.87	57.21	13.53	26.79	16.82
PGO-S2D	Source-only	50.60	53.79	18.51	22.81	16.69
PGO-S2D	Vanilla MT	51.86	54.98	19.77	23.11	17.96
PGO-S2D	PLAID	53.33	56.45	21.25	24.56	19.55
PGO-D2S	Source-only	33.88	42.78	6.26	19.44	11.99
PGO-D2S	Vanilla MT	39.09	44.19	9.28	24.04	16.09
PGO-D2S	PLAID	39.39	46.26	8.52	25.10	16.41

Table A3. Detection mAP by ground-truth action-segment duration d in seconds, averaged over tIoU thresholds 0.1 through 0.6. Entries report single-run results from the corresponding evaluated checkpoints. Duration boundaries are shared across all transfer settings and categories. N_{cls} counts represented classes in each bin. Classes with no ground truth in a bin are excluded from its mAP. During bin-specific evaluation, matching prioritizes in-bin annotations; predictions matched only to out-of-bin annotations are ignored, and unmatched predictions contribute false positives when their predicted duration belongs to the evaluated bin.

Setting	Duration (s)	N_{cls}	Source-Only	Vanilla MT	PLAID
PGO-S2L	$0 < d \leq 10$	3	12.22	13.33	14.41
PGO-S2L	$10 < d \leq 30$	5	14.86	16.55	17.02
PGO-S2L	$d > 30$	5	31.94	33.04	34.22
PGO-L2S	$0 < d \leq 10$	3	8.97	12.55	14.47
PGO-L2S	$10 < d \leq 30$	5	12.77	15.82	18.06
PGO-L2S	$d > 30$	5	29.92	34.05	36.00
PGO-S2D	$0 < d \leq 10$	4	22.44	23.75	25.31
PGO-S2D	$10 < d \leq 30$	5	27.16	28.42	29.93
PGO-S2D	$d > 30$	5	33.34	34.71	36.12
PGO-D2S	$0 < d \leq 10$	3	10.11	19.53	20.02
PGO-D2S	$10 < d \leq 30$	5	15.20	20.40	20.87
PGO-D2S	$d > 30$	5	25.40	27.48	27.62

Table A4. One-factor sensitivity comparison for the rank-gap coefficient and total cross-level borrowing budget. Other parameters use the settings in Section 4.1. Each entry is average mAP over tIoU thresholds 0.1 through 0.6. The default coefficient is 1 and the default borrowing budget is 16; the shared default configuration is evaluated once.

Parameter	Value	PGO-L2S	PGO-S2D
κ	0	29.58	33.21
κ	0.5	30.33	34.21
κ	1	30.64	35.23
κ	2	30.44	35.02
n_{borrow}	0	28.89	33.89
n_{borrow}	8	29.76	34.67
n_{borrow}	16	30.64	35.23
n_{borrow}	32	29.88	35.20

References

1. Zhang, C.L.; Wu, J.; Li, Y. ActionFormer: Localizing Moments of Actions with Transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.
2. Liu, Y.; Liu, Z.; Zhai, Y.; Li, W.; Doermann, D.; Yuan, J. STAT: Towards Generalizable Temporal Action Localization. *arXiv* **2024**, arXiv:2404.13311.
3. Pujol-Perich, D.; Clapés, A.; Escalera, S. SADA: Semantic Adversarial Unsupervised Domain Adaptation for Temporal Action Localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: Piscataway, NJ, USA, 2025; pp. 9219–9229.
4. Liang, J.; Hu, D.; Feng, J. Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. In Proceedings of the International Conference on Machine Learning (ICML), Virtually, 13–18 July 2020.
5. Tarvainen, A.; Valpola, H. Mean Teachers are Better Role Models: Weight-averaged Consistency Targets Improve Semi-supervised Deep Learning Results. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017.
6. Meng, L.; Ban, G.; Liu, F.; Qiu, W.; He, D.; Zhang, L.; Wang, S. Illegal Action Classification in Power Grid Operation Based on Cross-Domain Few-Shot Learning. *Power Syst. Big Data* **2024**, *27*, 69–76.
7. Meng, L.; He, D.; Ban, G.; Xi, G.; Li, A.; Zhu, X. Active Hard Sample Learning for Violation Action Recognition in Power Grid Operation. *Information* **2025**, *16*, 67. [CrossRef]
8. Meng, L.; He, D.; Ban, G.; Guo, S. Incremental Spatio-Temporal Augmented Sampling for Power Grid Operation Behavior Recognition. *Electronics* **2025**, *14*, 3579. [CrossRef]

9. Yan, S.; Yu, S. PowerOpsNet: Integrating YOLOv5-A and GCN for Real-Time Safety Behavior Recognition in Power Grid Environments. In *Proceedings of the 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*; IEEE: Piscataway, NJ, USA, 2025; pp. 6153–6159. [CrossRef]
10. Wu, T.; Wan, Y.; Wang, S.; Xiao, B.; Fang, C.; Li, P. A Video Abnormal Behavior Recognition Method for 10 kV Live Workers Based on Improved ST-GCN. *South. Power Syst. Technol.* **2024**, *18*, 31–39. 72. [CrossRef]
11. Yan, Y.; Zhou, Z.; Song, Y.; He, Y.; Wang, L.; Huang, H.; He, J.; Zhu, Z. Extensible Classification Method for Power Personnel Behavior Based on Pose Estimation. *Electr. Power* **2023**, *56*, 77–85. [CrossRef]
12. Ren, X.; Yang, Q.; Zhong, Y.; Yang, L.; Zhang, L.; Mao, X. Visual Prompt Method for Continual Learning in Ice Detection on Power Transmission and Distribution Lines. *Power Syst. Big Data* **2024**, *27*, 33–41.
13. Ban, G.; Fu, L.; Jiang, L.; Du, H.; Li, A.; He, Y.; Zhou, J. Two-Stage Dynamic Risk Identification of Complex Work Personnel Behavior Based on Image Screening. *Power Syst. Big Data* **2024**, *27*.
14. Liu, X.; Ban, G.; Wu, H.; Meng, L.; Zhou, J. A 3D Point Cloud Reconstruction Method Based on Improved ORBSLAM. *Power Syst. Big Data* **2024**, *27*, 72–79.
15. Ban, G.; Qiu, W.; Du, H.; Li, A.; Xi, G.; He, Y.; Zhou, J. Safe Zone Generation Technology in Complex Work Environment Based on Finite Element Method. *Power Syst. Big Data* **2024**, *27*, 54–63.
16. Qiu, W.; Liu, X.; Ban, G.; Fu, L.; Zhou, J. Neural Network-Based Recognition of Operational Features for Complex Operators. *Power Syst. Big Data* **2024**, *27*.
17. Cheng, L.-L.; Pei, Q.-G.; Yan, Y.-P.; Yu, Z.-Y.; Shao, Y.-Y.; Ruan, W.-C.; Yan, Q.; Lin, J.-X.; Cen, C.-Y.; Li, G.-G. Distribution Safety-Monitoring Data Cleaning Technology Based on OpenPose-GLCM-Hash. *Power Syst. Big Data* **2024**, *27*.
18. Lin, T.; Liu, X.; Li, X.; Ding, E.; Wen, S. BMN: Boundary-Matching Network for Temporal Action Proposal Generation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2019; pp. 3889–3898.
19. Xu, M.; Zhao, C.; Rojas, D.S.; Thabet, A.K.; Ghanem, B. G-TAD: Sub-Graph Localization for Temporal Action Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2020; pp. 10153–10162.
20. Tan, J.; Tang, J.; Wang, L.; Wu, G. Relaxed Transformer Decoders for Direct Action Proposal Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2021; pp. 13526–13535.
21. Liu, X.; Wang, Q.; Hu, Y.; Tang, X.; Zhang, S.; Bai, S.; Bai, X. End-to-End Temporal Action Detection with Transformer. *IEEE Trans. Image Process. (TIP)* **2022**, *31*, 5427–5441. [CrossRef] [PubMed]
22. Yi, F.; Wen, H.; Jiang, T. ASFormer: Transformer for Action Segmentation. In Proceedings of the British Machine Vision Conference (BMVC), Virtually, 22–25 November 2021.
23. Tang, T.N.; Kim, K.; Sohn, K. TemporalMaxer: Maximize Temporal Context with only Max Pooling for Temporal Action Localization. *arXiv* **2023**, arXiv:2303.09055.
24. Zhu, X.; Zhu, Y.; Chen, T.; Wu, W.; Dang, Y. FDDet: Frequency-Decoupling for Boundary Refinement in Temporal Action Detection. In *Advanced Intelligent Computing Technology and Applications*; Springer Nature: Singapore, 2025; pp. 205–217. [CrossRef]
25. Lu, H.; Yu, Y.; Lu, S.; Rajan, D.; Ng, B.P.; Kot, A.C.; Jiang, X. MambaTAD: When State-Space Models Meet Long-Range Temporal Action Detection. *IEEE Trans. Multimed.* **2026**, *28*, 8178–8192. [CrossRef]
26. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2017.
27. Diamant, I.; Rosenfeld, A.; Achituve, I.; Goldberger, J.; Netzer, A. De-Confusing Pseudo-Labels in Source-Free Domain Adaptation. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024, pp. 108–125. [CrossRef]
28. El-Sayed, A.; Torki, M. Zero-LEAD: Source-Free Universal Domain Adaptation for Abdominal Multi-Organ Segmentation. In *Proceedings of the 2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: Piscataway, NJ, USA, 2026; pp. 6715–6723. [CrossRef]
29. Xiao, Y.; Wei, P.; Wang, G.; Liu, C.; Lin, L. Robust Source-Free Domain Adaptation From Non-Robust Source Models. *IEEE Trans. Image Process.* **2026**, *35*, 2350–2363. [CrossRef] [PubMed]
30. Liu, C.; Lu, Y. ME-SFDA: Marginal Exploration with Pyramidal Atkinson-Shiffrin Memory for Source-Free Domain Adaptation. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 23658–23666. [CrossRef]
31. Peng, K.; Li, Y. PCMA: Prompt-guided Cross-domain Multi-prototype Alignment for Source-Free Domain Adaptation. In Proceedings of the 2026 International Conference on Multimedia Retrieval, Amsterdam, The Netherlands, 16–19 June 2026; pp. 1138–1146. [CrossRef]
32. Dadashzadeh, A.; Esmati, P.; Mirmehdi, M. Co-STAR: Collaborative Curriculum Self-Training with Adaptive Regularization for Source-Free Video Domain Adaptation. In *Proceedings of the 2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: Piscataway, NJ, USA, 2026; pp. 7821–7831. [CrossRef]

33. Wang, Y.; Gong, P.; Wu, M.; Ott, F.; Li, X.; Xie, L.; Chen, Z. Temporal Source Recovery for Time-Series Source-Free Unsupervised Domain Adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**, *48*, 9589–9605. [CrossRef] [PubMed]
34. Dai, Q.; Yang, S. Free on the Fly: Enhancing Flexibility in Test-Time Adaptation with Online EM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2025; pp. 9538–9548.
35. Mehrbod, P.; Vianna, P.; Nanfack, G.; Wolf, G.; Belilovsky, E. Test Time Adaptation Using Adaptive Quantile Recalibration. In *Proceedings of the 2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: Piscataway, NJ, USA, 2026; pp. 5986–5995. [CrossRef]
36. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the 37th International Conference on Machine Learning (ICML), 13–18 July 2020; Volume 119, pp. 1597–1607.
37. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2020.
38. Bardes, A.; Ponce, J.; LeCun, Y. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. In Proceedings of the International Conference on Learning Representations (ICLR), Online, 25–29 April 2022.
39. Hu, Y.; Yang, G.; Tan, Z.; Huang, X.; Huang, K.; Wang, Q.F. Covariance-Based Space Regularization for Few-Shot Class Incremental Learning. In Proceedings of the Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 26 February–6 March 2025; pp. 9548–9558.
40. Asaad, I.; Shadaydeh, M.; Denzler, J. Gradient Extrapolation for Debiased Representation Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2025; pp. 3819–3829.
41. Sepanj, M.H.; Ghojogh, B.; Moradi, S.; Fieguth, P. Kernel VICReg for Self-Supervised Learning in Reproducing Kernel Hilbert Space. *Big Data Cogn. Comput.* **2026**, *10*, 78. [CrossRef]
42. Roy, O.; Vetterli, M. The Effective Rank: A Measure of Effective Dimensionality. In Proceedings of the European Signal Processing Conference (EUSIPCO), Poznań, Poland, 3–7 September 2007.
43. Cho, Y.; Kim, Y.; Yoon, J.; Hong, S.; Lee, D. Feature Augmentation Based Test-Time Adaptation. In Proceedings of the Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 26 February–6 March 2025; pp. 6838–6847.
44. Yang, S.; Wang, Y.; van de Weijer, J.; Herranz, L.; Jui, S. Exploiting the Intrinsic Neighborhood Structure for Source-Free Domain Adaptation. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Online, 6–14 December 2021.
45. Yang, S.; Wang, Y.; Wang, K.; Jui, S.; van de Weijer, J. Attracting and Dispersing: A Simple Approach for Source-free Domain Adaptation. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Online, 28 November–9 December 2022.
46. He, Q.; Wu, X.; He, J.Y.; Li, S. Dual-Rate Dynamic Teacher for Source-Free Domain Adaptive Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2025; pp. 2067–2076. [CrossRef]
47. Idrees, H.; Zamir, A.R.; Jiang, Y.G.; Gorban, A.; Laptev, I.; Sukthankar, R.; Shah, M. The THUMOS challenge on action recognition for videos "in the wild". *Comput. Vis. Image Underst.* **2017**, *155*, 1–23. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.