Token-Aware Representation Augmentation for Fine-Grained Semi-Supervised Learning

Hongyang He[1†*], Yan Zhong[2*], Xinyuan Song[3*], Daizong Liu[4], Victor Sanchez[1]
[1]University of Warwick, [2]Peking University, [3]Emory University, [4]Wuhan University
[*]Equal contribution. [†]Corresponding Author.
Hongyang.He@warwick.ac.uk,zhongyan@stu.pku.edu.cn

FixMatch is a widely adopted semi-supervised learning (SSL) framework that relies on consistency regularization between weakly and strongly augmented versions of unlabeled data. In the case of image classification, its reliance on indiscriminate image-level augmentations often leads to overfitting on early confident predictions while neglecting semantically rich but underexplored features. In this work, we introduce Token-Aware FixMatch (TA-FixMatch), a novel SSL framework that operates at the token representation level to enhance feature diversity and generalization. Specifically, we propose a token-aware masking strategy that identifies and softly suppresses the most influential tokens contributing to high-confidence predictions; and a structured token-level augmentation pipeline that perturbs, reorganizes, and semantically enriches the remaining tokens. These representation-level augmentations guide the model to attend to alternative evidence and discover complementary features, which is particularly beneficial in fine-grained classification tasks. Extensive experiments on standard (CIFAR-100, STL-10) and fine-grained (CUB-200-2011, NABirds, Stanford Cars) benchmarks demonstrate that TA-FixMatch outperforms existing state-of-the-art SSL methods under low-label regimes.

1. Introduction

Semi-supervised learning (SSL) has emerged as a powerful paradigm for training high-performance models with limited labeled data by leveraging the vast availability of unlabeled samples. Among recent SSL methods [1–10], FixMatch has gained significant attention due to its simplicity and effectiveness. It combines pseudo-labeling with consistency regularization: given a weakly augmented unlabeled sample that yields a high-confidence prediction, FixMatch enforces the model to remain consistent under strong augmentation of the same sample [1]. This principle has achieved competitive results across a range of image classification tasks, including standard benchmarks like CIFAR-100 and STL. FixMatch has also inspired the development of many variants [2–4].

Despite its success, FixMatch suffers from a crucial limitation that is often overlooked; namely, its reliance on indiscriminate image-level augmentations. Popular strong augmentation strategies; i.e., RandAugment and CutOut, perturb pixel-level statistics, operating blindly over the entire image with no semantic awareness [11]. This not only risks removing discriminative content necessary for prediction but also reinforces the model's early confident patterns rather than encouraging exploration of underrepresented features. This limitation becomes particularly evident in fine-grained image classification, where subtle and spatially localized visual cues often determine class identity. In such scenarios, over-reliance on a few confident image regions can lead to feature collapse and hinder generalization [12].

To address these challenges, we propose TA-FixMatch, an SSL framework that shifts augmentation from the pixel domain to the internal token representation space. Our key insight is that modern vision backbones, e.g., CNNs and Vision Transformers, encode images into a sequence of token embeddings that capture semantically meaningful patterns. By identifying and manipulating these tokens directly, we can design stronger and more targeted augmentations to encourage learning of a richer set of features. TA-FixMatch introduces two complementary components: (1) a token-aware masking mechanism that evaluates token importance via gradient-based saliency and softly

Second Conference on Parsimony and Learning (CPAL 2026).

suppresses the most influential tokens associated with confident predictions; and (2) a structured token-level augmentation pipeline that enhances the diversity of the remaining token set through perturbation, local shuffling, and semantic injection. These operations encourage redistributing attention to less confident but potentially informative regions, mitigating overfitting and enhancing generalization.

We conduct extensive experiments on standard and fine-grained image classification benchmarks. TA-FixMatch achieves state-of-the-art performance under the 1% and 5% label regimes on challenging datasets; i.e., CUB-200-2011, NABirds, and Stanford Cars; demonstrating its capability to discover and exploit subtle features often overlooked.

2. Related works

SSL has gained importance in deep learning due to its capability to exploit unlabeled data to improve performance. Early SSL methods rely on pseudo-labeling [5] and consistency regularization [6]. Among them, FixMatch has emerged as a seminal work by combining high-confidence pseudo-label selection on weakly augmented samples with consistency regularization on strongly augmented samples [1, 13, 14]. Its simplicity and effectiveness have spurred a series of follow-up works aimed at improving either the pseudo-labeling mechanism or the augmentation strategies; e.g., FlexMatch [2], FreeMatch [3], Dash [7], and SoftMatch [4]. These methods leverage adaptive confidence thresholds or curriculum-style pseudo-labeling to improve the selection of unlabeled samples. Despite their successes, most of them retain the core design of FixMatch; i.e., strong image-level augmentations, which may suppress or distort important semantic cues, particularly in fine-grained image classification where discriminative features are often subtle and spatially localized [15, 16].

Recent advances have addressed the aforementioned challenges by proposing pseudo-labeling strategies tailored for fine-grained visual classification (FGVC). For example, PEPL [17] incorporates class activation maps into a two-stage pseudo-labeling process, refining label generation and mixing to preserve subtle discriminative features. SoC [18], on the other hand, relaxes the hard-label assumption by jointly optimizing soft-label expansion and shrinkage, guided by class transition tracking to adaptively group visually similar classes. Both methods demonstrate significant gains under low-label regimes, underscoring the importance of semantic-aware and noise-resilient pseudo-labeling in semi-supervised FGVC. From a feature learning perspective, recent works explore how neural networks selectively attend to high-confidence image regions, potentially neglecting semantically valuable but less confident regions. This phenomenon becomes more pronounced in FGVC [11, 19], where the model's over-reliance on dominant features can lead to suboptimal generalization. Motivated by these insights [20], we propose to shift the augmentation from pixels to internal token representations, offering a new perspective on augmenting and regularizing model predictions in SSL.

3. Preliminaries: Revisiting FixMatch and Its Limitations

FixMatch is a widely adopted framework for SSL. It trains a classifier using both labeled data and pseudo-labeled data that are initially unlabeled [1]. The core of FixMatch consists of two components: (1) cross-entropy loss on weakly augmented labeled samples and (2) consistency regularization that enforces a strongly augmented view of an unlabeled sample to match its pseudo-label generated from a weakly augmented view [21]. Formally, let $x_l \in \mathcal{D}_l$ be the labeled data, $x_u \in \mathcal{D}_u$ the unlabeled data, $F(\cdot)$ the model's prediction function, and $p_i(F(\cdot))$ the softmax probability for class i. FixMatch generates a pseudo-label $\hat{y}_u = \arg\max_i p_i(F(\alpha(x_u)))$ for unlabeled sample x_u if the model's prediction exceeds a confidence threshold τ, where $\alpha(\cdot)$ denotes weak augmentation. It then enforces prediction consistency on the strongly augmented input $A(x_u)$:

$$\mathcal{L}_u = \mathbb{E}_{x_u \sim \mathcal{D}_u} \left[\mathbb{I} \left(\max p_i(F(\alpha(x_u))) \geq \tau \right) \cdot \mathrm{CE}(F(A(x_u)), \hat{y}_u) \right], \tag{1}$$

where $A(\cdot)$ denotes strong augmentation, and $\mathrm{CE}(\cdot, \cdot)$ is the cross-entropy loss. Although effective, the generalization capability of FixMatch can be limited by the randomness and coarseness of strong augmentations such as CutOut [22] or RandAugment [23]. These augmentations operate indiscriminately over image regions, often failing to reliably suppress features that have already been

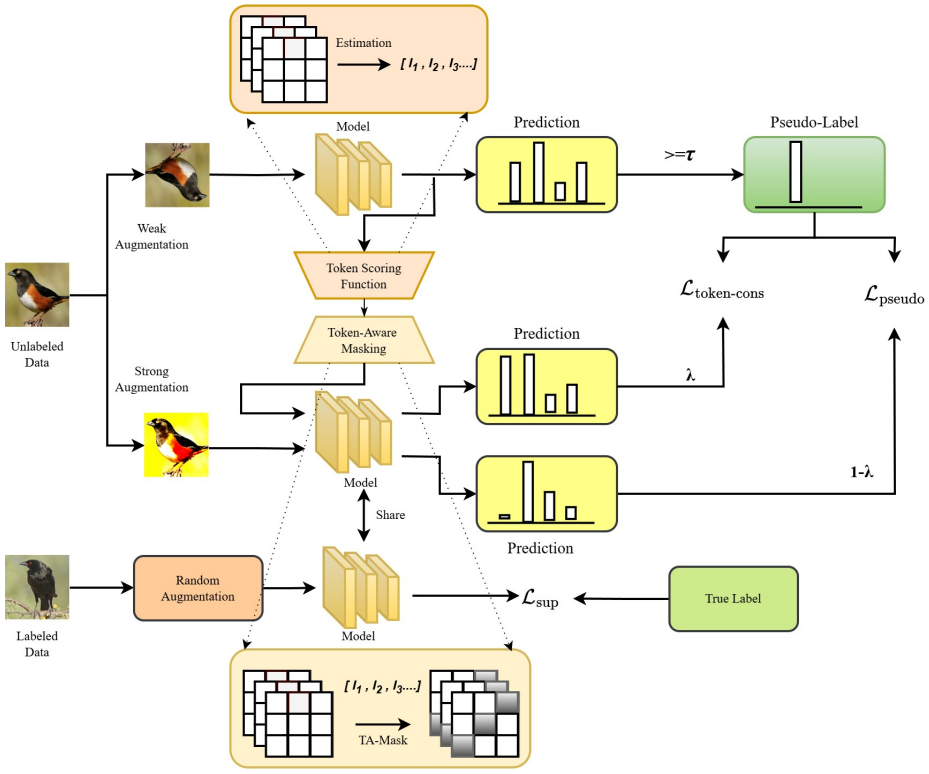


Figure 1: TA-FixMatch. The model generates pseudo-labels from weakly augmented unlabeled inputs, applies token-aware masking and augmentation to strong views, and optimizes a dual-path loss combining token-level consistency and pseudo-label supervision.

learned by the model. Consequently, the consistency objective may not sufficiently guide the model to learn novel [24], discriminative features that are underrepresented in early training phases. This observation motivates us to rethink the design of the strong augmentation $A(\cdot)$ and explore a more targeted augmentation strategy that interacts directly with the model's internal token representations, enabling feature-level supervision beyond the pixel space [11].

4. Proposed Token-Aware FixMatch

TA-FixMatch improves FixMatch by introducing a representation-level augmentation mechanism. The key idea is to identify and mask tokens (patch-level embeddings) that are highly responsible for the model's confident predictions, thus encouraging the model to learn from alternative tokens and acquire a more comprehensive understanding of the input space. TA-FixMatch consists of two main components: Token Scoring Function and Token-Aware Masking (see Figure 1).

4.1. Token Scoring Function

Let $x_u \in \mathcal{D}_u$ denote an unlabeled input and $X_u \in \mathbb{R}^{P\times d}$ its tokenized representation extracted by a backbone (e.g., patch embeddings from a CNN or a ViT), where P is the number of tokens and d is the token dimension. Each token $x_p \in \mathbb{R}^d$ corresponds to the p-th row of X_u. For an unlabeled input that has been assigned a pseudo-label with prediction $\hat{y}_u$, we evaluate the importance of each token x_p toward the predicted class:

$$I_p = \left\| \frac{\partial \mathcal{L}_{CE}(F(X_u), \hat{y}_u)}{\partial x_p} \right\|_2, \tag{2}$$

where $\mathcal{L}_{CE}$ is the standard cross-entropy loss. The gradient-based norm $\|\cdot\|_2$ reflects how sensitive the model's output is to perturbations at each token position.

4.2. Token-Aware Masking

Given the importance scores $\{I_p\}_{p=1}^{P}$ as computed based on gradient importance (see Eq. 2), we perform soft-masking on the top-k most influential tokens by attenuating their embeddings based

on their relative importance. Let $k = \lfloor \rho \cdot P \rfloor$ denote a sample-specific number of tokens, where $\rho \in (0, 1)$ is a predefined masking ratio. In this work, we set $\rho = 0.1$, i.e., the top 10% tokens are selected per sample. For the selected tokens $\{x_p\}$, we apply a multiplicative masking strategy using a decay coefficient $\beta_p \in (0, 1)$:

$$x_p \leftarrow \beta_p \cdot x_p, \quad \forall p \in \text{Top-}k(\{I_p\}), \tag{3}$$

where $\beta_p = \exp\left(-\mu \cdot \frac{I_p}{\max\left(\{I_p\}\right)}\right)$ is a soft suppression factor controlled by a scaling parameter μ. Larger values of μ result in more aggressive down-weighting of important tokens; we set $\mu = 1.5$ by default in practice. This exponential formulation ensures that more important tokens are suppressed more strongly, while less dominant tokens are preserved.

The resulting softly masked token sequence, $\tilde{X}_u$, is then passed to the model as the strongly augmented input. This augmentation strategy gently suppresses the highly confident internal representations without completely erasing them, thereby encouraging the model to redistribute its attention to less exploited parts of the input space and to learn more diverse, decision-supportive features.

Token-Level Augmentation. To enhance the model's capability to leverage underexplored semantics in unlabeled data, we propose a structured token-level augmentation framework that complements token-aware masking in TA-FixMatch. Rather than applying independent perturbations arbitrarily, our method treats each token as a unit of semantic reasoning and designs a progressive augmentation process that expands the model's representational space. The goal is not only to introduce diversity but to explicitly encourage the model to attend to alternative or secondary evidence within each sample.

Given a tokenized representation $X_u \in \mathbb{R}^{P \times d}$ from an unlabeled input with confident pseudo-label $\hat{y}_u$, we first apply token-aware masking to suppress dominant evidence based on gradient importance, as explained before. The remaining tokens, potentially containing complementary semantics, are then subjected to a three-step augmentation process designed to perturb them, re-organize them, and apply semantic injection.

In the first step, we apply contextual perturbation by injecting adaptive noise into individual token embeddings. This noise can take either a stochastic or adversarial:

$$x_p \leftarrow x_p + \epsilon \cdot \frac{g_p}{\|g_p\|_2}, \quad \text{where} \quad g_p = \frac{\partial \mathcal{L}_{CE}(F(X_u), \hat{y}_u)}{\partial x_p}, \quad p \notin \text{Top-}k(\{I_p\}). \tag{4}$$

where $\epsilon = 0.1$ controls the perturbation strength. This operation amplifies latent sensitivity of subdominant tokens, allowing the model to discover new decision-relevant patterns.

In the second step, we perform semantic re-organization by altering the structural arrangement of local token groups. This involves lightweight shuffling or position-aware mixing within short token windows to break positional bias while maintaining content continuity:

$$x_{p=1}, x_{p=2}, \cdots \leftarrow \text{LocalShuffle}(x_{p=1}, x_{p=2}, \dots), \quad p \in W \subset [1, P], \tag{5}$$

where W denotes a fixed-size sliding window (e.g., size 3–5) over token positions for local mixing. This allows the model to reinterpret tokens in different local contexts, improving its robustness to spatial deformation and redundancy.

The final step introduces semantic injection, where a small fraction of low-importance tokens are replaced with embeddings from a class-agnostic dictionary:

$$x_p \leftarrow \mathcal{E}(j_p), \quad j_p \sim \mathcal{U}(1, |\mathcal{E}|), \quad p \in S, \tag{6}$$

where $S = \{p \in [1, P] \mid I_p \leq \gamma\}$ defines the set of low-importance tokens eligible for semantic injection, and γ is a threshold empirically chosen based on the distribution of importance scores. Here, j_p denotes the sampled index for each token p, independently drawn from the uniform distribution over the dictionary entries. $\mathcal{E} : \{1, \dots, |\mathcal{E}|\} \rightarrow \mathbb{R}^d$ denotes a class-agnostic embedding dictionary. In practice, we construct the class-agnostic dictionary $\mathcal{E}$ by averaging randomly selected small groups

of token embeddings from unlabeled data, which provides neutral yet semantically coherent vectors for semantic injection. This step broadens the token distribution encountered during training, increasing the diversity of support for the pseudo-label decision.

Through these three steps—perturbation, re-organization, and semantic injection—our token-level augmentation pipeline systematically expands the model's view of unlabeled data. Each operation is designed not for generic regularization, but to reinforce the learning of complementary features that are often overlooked by confidence-based selection. The resulting strongly augmented tokens $\tilde{X}_u$ are passed through the model for consistency training against the original pseudo-label $\hat{y}_u$. This strategy aligns well with the philosophy of TA-FixMatch; namely, to guide the model away from overfitting to early confident patterns and towards a more holistic understanding of the input.

4.3. Objective Function

The overall optimization objective of TA-FixMatch combines supervised learning on labeled data with two complementary losses over pseudo-labeled data that are originally unlabeled: a consistency loss on perturbed tokens and a pseudo-label supervision loss on the original token representation. To better balance the two learning objectives, we introduce a trade-off parameter $\lambda \in [0, 1]$, which controls the relative contribution of each loss term. The total objective is defined as:

$$\mathcal{L}_{\text{TA-FixMatch}} = \mathcal{L}_{\text{sup}} + \lambda \mathcal{L}_{\text{token-cons}} + (1 - \lambda)\mathcal{L}_{\text{pseudo}}. \tag{7}$$

The supervised loss is computed over labeled samples with weak augmentation. Let $(x_l, y_l) \sim \mathcal{D}_l$ denote a labeled sample and its ground-truth label:

$$\mathcal{L}_{\text{sup}} = \mathbb{E}_{(x_l, y_l)\sim\mathcal{D}_l} \left[\text{CE}(F(\alpha(x_l)), y_l)\right]. \tag{8}$$

For unlabeled data, we first generate a pseudo-label $\hat{y}_u = \arg\max_i p_i(F(\alpha(x_u)))$ for unlabeled sample $x_u \sim X_u$ if the model's confidence exceeds a threshold τ. We then apply token-aware masking followed by token-level augmentation to obtain a strongly perturbed representation $\tilde{X}_u = \text{TokenAug}(\text{Mask}(X_u))$, where $\text{Mask}(\cdot)$ and $\text{TokenAug}(\cdot)$ are the functions implementing token-aware masking and token-level augmentation, respectively, as explained in Section 4.

The token-consistency loss encourages the model to make stable predictions on perturbed token sequences:

$$\mathcal{L}_{\text{token-cons}} = \mathbb{E}_{x_u\sim\mathcal{D}_u} \left[\mathbb{I}\left(\max p_i(F(\alpha(x_u))) \geq \tau\right) \cdot \text{CE}(F(\tilde{X}_u), \hat{y}_u)\right]. \tag{9}$$

In parallel, the pseudo-label supervision loss encourages the model to predict consistently on the original, unmasked token sequence:

$$\mathcal{L}_{\text{pseudo}} = \mathbb{E}_{x_u\sim\mathcal{D}_u} \left[\mathbb{I}\left(\max p_i(F(\alpha(x_u))) \geq \tau\right) \cdot \text{CE}(F(X_u), \hat{y}_u)\right]. \tag{10}$$

Here, $p_i(F(\cdot))$ denotes the softmax probability of class i, and $\mathbb{I}(\cdot)$ is the confidence threshold indicator. Together, $\mathcal{L}_{\text{token-cons}}$ and $\mathcal{L}_{\text{pseudo}}$ form a dual-path supervision scheme, with λ providing explicit control over the relative emphasis on perturbed versus original token consistency (see Figure 1). This flexibility allows TA-FixMatch to dynamically trade off between exploration of alternative features and consolidation of confident patterns, leading to better generalization and more robust feature representations.

5. Theoretical Analysis

To understand the effect of token-aware masking in TA-FixMatch, we examine how the decay coefficient β_p (see Eq. 3) applied to influential tokens changes the gradient signal used for updating token embeddings. Since the consistency loss is computed on the masked representation $\tilde{X}_u$, it is important to quantify how much the gradient at each token can shrink after masking. The following theorem provides a formal upper bound on this change and shows that soft masking reduces the gradient in a controlled and stable manner.

Theorem 1 (Soft Mask Gradient Decay Property)**.** *Let $F : \mathbb{R}^{P\times d} \to \mathbb{R}^C$ be a differentiable prediction function, and let the input token sequence be $X_u = [x_1, \ldots, x_P] \in \mathbb{R}^{P\times d}$, where each token $x_p \in \mathbb{R}^d$*

denotes the p-th row of X_u. For the pseudo-label $\hat{y}_u = \arg\max_i p_i(F(X_u))$, define the token-wise importance $I_p = \left\| \frac{\partial \mathcal{L}_{CE}(F(X_u), \hat{y}_u)}{\partial x_p} \right\|_2$. Consider the soft-masked tokens $\tilde{x}_p = \beta_p x_p$ with

$$\beta_p = \exp\left(-\mu \cdot \frac{I_p}{\max\left(\{I_p\}\right)} \right), \quad \mu > 0, \tag{11}$$

and let $\tilde{X}_u = [\tilde{x}_1, \ldots, \tilde{x}_P]$. Suppose the Hessian with respect to x_p satisfies $\left\| \frac{\partial^2 \mathcal{L}_{CE}(F(X_u), \hat{y}_u)}{\partial x_p^2} \right\|_2 \leq H$ for all X on the line segment between X_u and $\tilde{X}_u$, and define $M = H\|x_p\|_2$. Then the gradient norm after masking satisfies

$$\left\| \frac{\partial \mathcal{L}_{CE}(F(\tilde{X}_u), \hat{y}_u)}{\partial x_p} \right\|_2 \leq \beta_p \left\| \frac{\partial \mathcal{L}_{CE}(F(X_u), \hat{y}_u)}{\partial x_p} \right\|_2 + (1 - \beta_p) M. \tag{12}$$

This result formalizes the effect of soft masking in TA-FixMatch. The bound shows that the gradient at each masked token decays in proportion to the decay coefficient β_p, while the residual term is controlled by the curvature of the loss landscape. Consequently, tokens with large importance scores receive a strictly reduced gradient after masking, which pushes the model to rely more on alternative tokens during consistency training. The proof of Theorem 1 is provided in Section A.

To complement the gradient decay analysis in Theorem 1, we now study how robust the token-consistency branch is under token-level augmentation. In TA-FixMatch, the original representation X_u is transformed into $\tilde{X}_u$ through perturbation, local re-organization, and semantic injection, and the consistency loss $\mathcal{L}_{\text{token-cons}}$ is optimized on $F(\tilde{X}_u)$. A natural question is whether these perturbations can destabilize the training dynamics by producing large gradients with respect to the original input. Theorem 2 answers this question by providing an explicit upper bound on $\|\nabla_{X_u} \mathcal{L}_{\text{token-cons}}\|_F$ in terms of $\|\nabla X_u \mathcal{L}_{\text{pseudo}}\|_F$ and the perturbation scale δ of the augmentation operator T.

Theorem 2 (Consistency Loss Gradient Stability Property)**.** *Let $F : \mathbb{R}^{P \times d} \to \mathbb{R}^C$ be Lipschitz continuous and twice differentiable, and let $X_u \in \mathbb{R}^{P \times d}$ be an unlabeled token sequence with pseudo-label $\hat{y}_u$ generated from the weak view with confidence greater than τ. Define the token-consistency loss $\mathcal{L}_{\textit{token-cons}} = CE(F(\tilde{X}_u), \hat{y}_u)$, where $\tilde{X}_u$ is obtained from X_u through a token-level augmentation transformation T satisfying*

$$\|\tilde{X}_u - X_u\|_F \leq \delta. \tag{13}$$

Assume the gradient of F has Lipschitz constant L and the Hessian satisfies

$$\left\| \frac{\partial^2 \mathcal{L}_{CE}(F(X_u), \hat{y}_u)}{\partial X^2} \right\|_2 \leq H_{\max} \tag{14}$$

for all X on the line segment between X_u and $\tilde{X}_u$. Then the gradient of the consistency loss with respect to the original input satisfies

$$\|\nabla_{X_u} \mathcal{L}_{\textit{token-cons}}\|_F \leq \left\|\nabla_{X_u} \mathcal{L}_{\textit{pseudo}}\right\|_F + \delta \left(L + H_{\max} \cdot \|X_u\|_F\right), \tag{15}$$

where $\mathcal{L}_{\textit{pseudo}} = CE(F(X_u), \hat{y}_u)$ is the pseudo-label loss computed on the original token sequence.

Theorem 2 shows that, as long as the token-level augmentation satisfies $\|\tilde{X}_u - X_u\|_F \leq \delta$ and the model has bounded first- and second-order behavior, the gradient of the token-consistency loss with respect to X_u remains close to that of the pseudo-label loss. The deviation is controlled by the term $\delta\left(L + H_{\max}\|X_u\|_F\right)$, which links the strength of augmentation to the smoothness of the model. Consequently, the token-augmented branch of TA-FixMatch can introduce diverse and structured perturbations without causing unstable or exploding gradients. Together with Theorem 1, this provides a theoretical justification that TA-FixMatch suppresses dominant tokens and leverages alternative evidence in a stable manner. A proof of Theorem 2 is presented in Section B.

6. Evaluation and Analysis

Datasets. We conduct classification experiments on several image datasets to evaluate the generalization performance of TA-FixMatch under limited supervision. Specifically, we consider standard

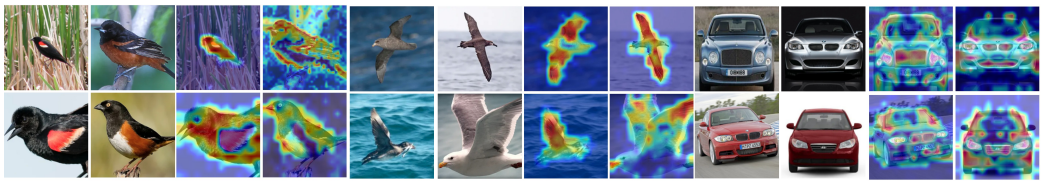


Figure 2: Visualizations generated by TA-FixMatch on CUB-200-2011 and Stanford Car.

Table 1: Accuracy (% $\pm$ standard deviation) Supervised Learning (SL), FixMatch, and TA-FixMatch under 1%, 5%, and 10% labeled data settings on CIFAR-100 and STL-10.

Method	**CIFAR-100**			**STL-10**		
	1%	5%	10%	1%	5%	10%
SL [1]	12.03 $\pm$ 0.25	41.12 $\pm$ 0.38	62.91 $\pm$ 0.41	24.58 $\pm$ 1.55	39.77 $\pm$ 1.00	65.12 $\pm$ 0.60
FixMatch [1]	54.66 $\pm$ 0.59	70.88 $\pm$ 0.52	**76.95** $\pm$ 0.20	68.95 $\pm$ 3.88	88.22 $\pm$ 0.81	92.87 $\pm$ 0.22
TA-FixMatch (Ours)	**55.87** $\pm$ 0.47	**71.43** $\pm$ 0.36	76.91 $\pm$ 0.26	**70.41** $\pm$ 4.01	**89.14** $\pm$ 0.97	**93.29** $\pm$ 0.21

benchmarks; i.e., CIFAR-100 [25] and STL-10 [26]; as well as fine-grained datasets; i.e., CUB-200-2011 [27], Stanford Car [28] and NABirds [29]. Across all datasets, we follow the SSL protocol, using only 1%, 5%, and 10% of the training data as labeled samples while treating the rest as unlabeled data. Each experiment is repeated five times with different random seeds. We report the mean classification accuracy along with standard deviation to ensure robustness and reproducibility.

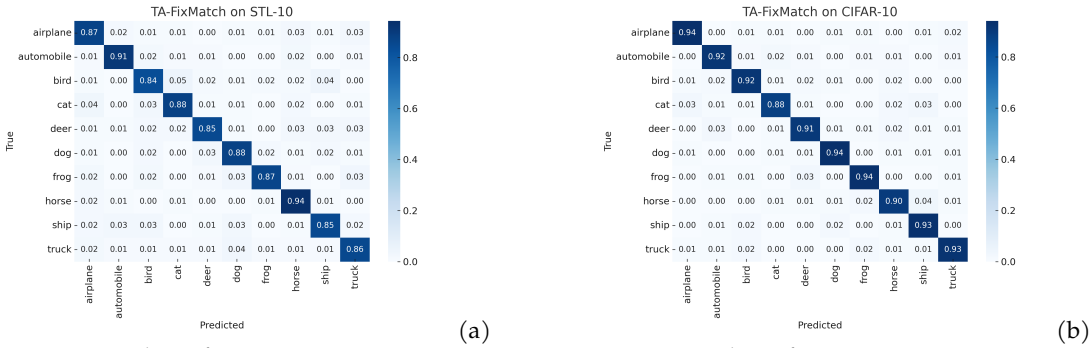
Implementation details. We adopt the WRN-28-8 [30] architecture for all datasets. Our implementation is based on PyTorch and closely follows the standard FixMatch pipeline. Specifically, we apply weak augmentation consisting of random horizontal flipping and random cropping, and strong augmentation using RandAugment. We use confidence threshold $\tau = 0.95$ for pseudo-labeling and SGD with momentum for optimization. The batch size is set to 64 for labeled data and 320 for unlabeled data, with a learning rate of 0.03 and a cosine learning rate schedule. The number of training steps is dataset-dependent [11]. To balance dual-path supervision, we set the trade-off parameter $\lambda = 0.65$ in all experiments (see Eq. 7).

Results. As shown in Figure 2, TA-FixMatch consistently highlights multiple semantically relevant regions within each object. On CUB-200-2011, the model attends to both primary (e.g., wings, tail) and secondary (e.g., head, belly) features of birds, indicating its capability to discover fine-grained cues beyond dominant parts. Similarly, on Stanford Car, attention is distributed across headlights, grilles, and license plates, suggesting that token-level augmentation leads to a more holistic understanding of structured objects.

As shown in Table 1, TA-FixMatch consistently surpasses FixMatch, especially under 1% labeling. On CIFAR-100 and STL-10, TA-FixMatch reaches 55.87% and a 1.46% gain, exceeding FixMatch by 1.21% and 1.46%, respectively. These improvements derive from selectively suppressing overconfident tokens while augmenting under-explored ones, rather than relying solely on image-level perturbations. TA-FixMatch also remains competitive at 10% labels, indicating scalability. Table 2 further confirms superiority on fine-grained benchmarks: with 1% labels, TA-FixMatch attains 19.01% on CUB-200-2011 and 17.08% on NABirds, improving over PEPL by 0.75% and 0.93%. Although PEPL leads in certain 10% settings, TA-FixMatch stays competitive across regimes, demonstrating robustness in fine-grained scenarios where spatially localized cues affect class separability.

Table 2: Accuracy (%) of several SOTA SSL methods (all based on ResNet backbone) under 1%, 5%, and 10% labeled data five times on fine-grained datasets.

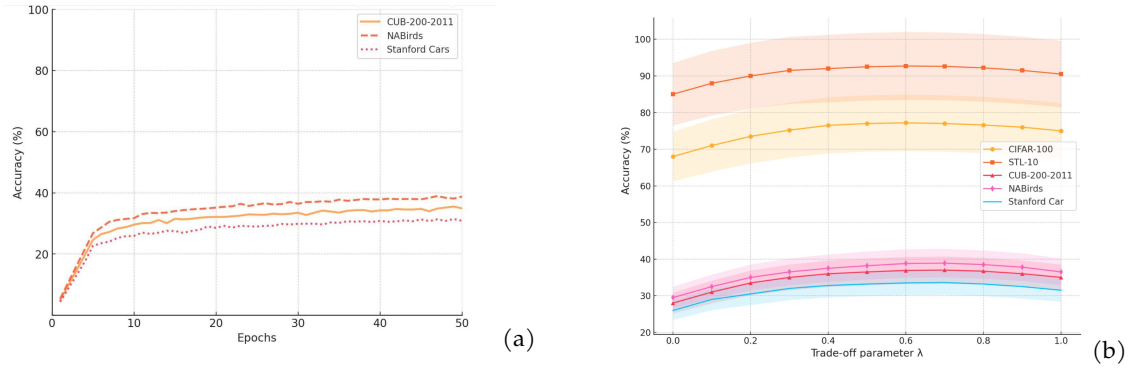
Method	**CUB_200_2011**			**NABirds**			**Stanford Car**		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
SL	12.20 $\pm$0.21	25.35 $\pm$ 0.18	28.61 $\pm$ 0.24	11.60 $\pm$ 0.17	24.40 $\pm$ 0.20	27.70 $\pm$ 0.23	10.42 $\pm$ 0.19	21.66 $\pm$ 0.22	24.54 $\pm$ 0.20
Pi-Model [31]	11.58 $\pm$ 0.24	23.79 $\pm$ 0.19	25.52 $\pm$ 0.28	10.85 $\pm$ 0.22	22.15 $\pm$ 0.25	24.15 $\pm$ 0.27	9.40 $\pm$ 0.23	20.10 $\pm$ 0.24	23.01 $\pm$ 0.26
Pseudo-Label [5]	14.95 $\pm$ 0.26	30.11 $\pm$ 0.22	32.71 $\pm$ 0.25	13.20 $\pm$ 0.23	28.80 $\pm$ 0.31	31.10 $\pm$ 0.29	12.38 $\pm$ 0.28	26.45 $\pm$ 0.27	26.12 $\pm$ 0.30
FlexMatch [2]	15.30 $\pm$ 0.29	31.45 $\pm$ 0.27	30.61 $\pm$ 0.24	13.90 $\pm$ 0.25	30.00 $\pm$ 0.28	32.80 $\pm$ 0.32	13.95 $\pm$ 0.27	26.90 $\pm$ 0.26	26.70 $\pm$ 0.31
FixMatch [1]	16.42 $\pm$ 0.30	33.88 $\pm$ 0.25	30.78 $\pm$ 0.31	14.20 $\pm$ 0.27	31.52 $\pm$ 0.29	33.15 $\pm$ 0.26	14.30 $\pm$ 0.28	26.85 $\pm$ 0.25	26.10 $\pm$ 0.33
PEPL [17]	18.26 $\pm$ 0.18	36.20 $\pm$ 0.21	**38.53** $\pm$ 0.20	16.15 $\pm$ 0.19	35.45 $\pm$ 0.22	37.00 $\pm$ 0.24	16.52 $\pm$ 0.21	30.30 $\pm$ 0.19	**32.72** $\pm$ 0.23
TA-FixMatch	**19.01** $\pm$ 0.17	**37.52** $\pm$ 0.19	35.02 $\pm$ 0.22	**17.08** $\pm$ 0.16	**36.90** $\pm$ 0.20	**38.45** $\pm$ 0.21	**17.30** $\pm$ 0.18	**30.45** $\pm$ 0.17	31.21 $\pm$ 0.20



(a) TA-FixMatch confusion matrix on STL-10.

(b) TA-FixMatch confusion matrix on CIFAR-10.

Figure 3: Confusion matrices of TA-FixMatch on (a) STL-10 and (b) CIFAR-10 under the 10% labeled data setting. TA-FixMatch produces strong diagonal patterns with low cross-class confusion, demonstrating robust pseudo-label quality and stable token-level consistency.



(a) (b)

Figure 4: (a) Accuracy curves over training epochs across several datasets under only 10% labeled data. (b) Performance variation with respect to the trade-off parameter λ under only 10% labeled data, where most datasets achieve peak accuracy around $\lambda = 0.65$.

As illustrated in Figure 3, TA-FixMatch exhibits clear diagonal confusion patterns on both STL-10 and CIFAR-10, indicating accurate predictions with minimal cross-class errors. This aligns with the gains in Table 1, where TA-FixMatch surpasses FixMatch under low-label settings. The reduced off-diagonal confusion shows that token-aware masking and token-level augmentation improve pseudo-label reliability and encourage the model to exploit complementary features, leading to more stable decision boundaries and better generalization. Figure 4 (a) illustrates the training accuracy curves of TA-FixMatch across multiple fine-grained datasets under the 10% labeled data setting. We observe that the model steadily improves throughout training on all datasets, indicating stable convergence behavior.

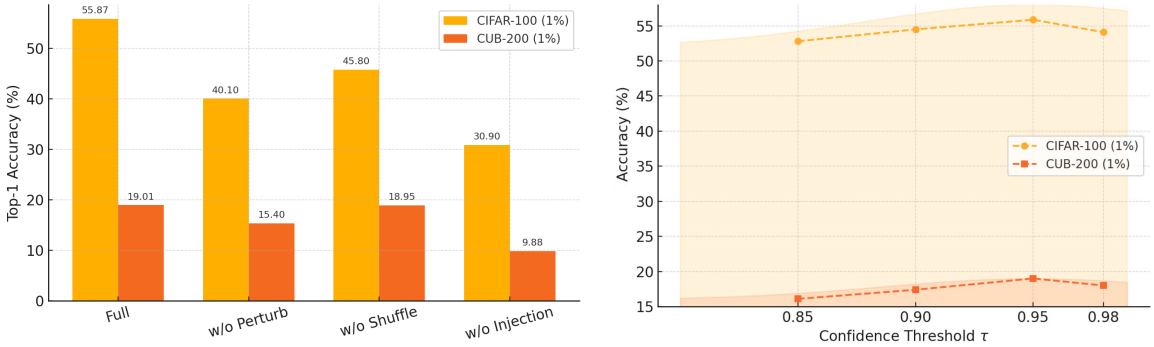
7. Ablation study

As shown in Table 3, both token-aware masking and token-level augmentation contribute significantly to the performance of TA-FixMatch. Removing either component results in a consistent drop in accuracy across all label settings. In particular, the absence of token-aware masking leads to a 2.05% drop on CIFAR-100 and a 2.11% drop on CUB-200 under the 1% labeled data regime. Token-level augmentation also provides noticeable gains, especially on fine-grained datasets, by enhancing semantic diversity. These results validate the complementary roles of both modules in improving generalization under limited supervision.

Figure 4 (b) depicts the impact of varying the trade-off parameter, λ, on accuracy under the 10% labeled data setting. Most datasets achieve peak performance when λ is 0.6-0.7, demonstrating the robustness of our chosen default configuration. However, slight deviations are observed, with CUB-200-2011, NABirds, and Stanford Car benefiting most significantly at exactly $\lambda = 0.65$, highlighting the importance of balancing token-consistency and pseudo-label supervision appropriately.

Table 3: Accuracy (%) on CIFAR-100 and CUB_200_2011 under 1%, 5%, and 10% labeled data settings when token-aware masking and token-level augmentation are removed.

Method	CIFAR-100			CUB_200_2011		
	1%	5%	10%	1%	5%	10%
w/o Token-Aware Masking	53.14 ± 0.18	70.65 ± 0.21	76.02 ± 0.23	16.82 ± 0.14	34.40 ± 0.19	33.25 ± 0.17
w/o Token-Level Augmentation	54.02 ± 0.20	70.88 ± 0.22	76.34 ± 0.24	17.45 ± 0.16	35.26 ± 0.18	33.88 ± 0.20
TA-FixMatch (full)	**55.87** ± 0.47	**71.43** ± 0.36	**76.91** ± 0.26	**19.01** ± 0.17	**37.52** ± 0.19	**35.02** ± 0.22



(a) (b)

Figure 5: (a) Impact of removing each token-level augmentation stage. (b) Performance with respect to confidence threshold τ, showing peak performance when $\tau = 0.95$.

Table 4: Accuracy (%) on CIFAR-100, CUB_200_2011, and NABirds using 1%, 5%, and 10% labeled data for different masking ratio values, ρ.

ρ	CIFAR-100			CUB_200_2011			NABirds		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.00	52.63 ± 0.19	65.26 ± 0.22	72.09 ± 0.24	16.52 ± 0.18	25.41 ± 0.21	32.23 ± 0.25	15.10 ± 0.20	24.32 ± 0.23	30.88 ± 0.26
0.05	52.43 ± 0.18	64.49 ± 0.21	**77.24** ± 0.25	15.16 ± 0.20	24.79 ± 0.19	31.36 ± 0.24	16.40 ± 0.21	24.95 ± 0.22	30.95 ± 0.27
0.10	**55.87** ± 0.47	**71.43** ± 0.36	76.91 ± 0.26	**19.01** ± 0.17	**37.52** ± 0.19	**35.02** ± 0.22	**17.08** ± 0.16	**36.90** ± 0.20	**38.45** ± 0.21
0.15	52.81 ± 0.20	65.24 ± 0.23	71.58 ± 0.25	16.25 ± 0.18	25.10 ± 0.20	32.77 ± 0.24	16.60 ± 0.22	25.25 ± 0.24	31.21 ± 0.28
0.20	43.48 ± 0.22	64.13 ± 0.26	60.29 ± 0.28	12.96 ± 0.20	24.02 ± 0.22	31.50 ± 0.26	14.75 ± 0.23	20.88 ± 0.25	26.15 ± 0.30

Figure 5 (a) shows that removing semantic injection causes the largest degradation, confirming its central role in driving feature diversity, while perturbation and shuffle yield smaller but consistent gains, indicating the three stages are complementary. Figure 5 (b) further shows that performance peaks at $\tau = 0.95$; both lower and higher thresholds harm pseudo-label reliability or coverage, supporting $\tau = 0.95$ as a stable default. Table 4 indicates that $\rho = 0.10$ provides the best overall results across datasets and label regimes, balancing confidence suppression and information retention. Although $\rho = 0.05$ occasionally performs well (e.g., 77.24% on CIFAR-100 with 10% labels), $\rho = 0.20$ consistently causes severe drops, confirming over-masking risk and validating $\rho = 0.10$ as a reliable default.

8. Conclusions

In this work, we proposed TA-FixMatch, a token-level augmentation framework that addresses the over-reliance on dominant features in SSL. By identifying and softly masking high-importance tokens, and introducing structured perturbations to less confident ones, TA-FixMatch is encouraged to discover complementary features often overlooked by standard approaches. Extensive experiments on standard and fine-grained benchmarks demonstrate that TA-FixMatch consistently improves generalization under limited supervision. Our findings suggest that controlling representation dynamics, rather than relying solely on pixel-space perturbations, can lead to more robust and semantically diverse SSL models.

References

[1] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[2] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in neural information processing systems*, 34:18408–18419, 2021.

[3] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022.

[4] Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*, 2023.

[5] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta, 2013.

[6] Erik Englesson and Hossein Azizpour. Consistency regularization can improve robustness to label noise. *arXiv preprint arXiv:2110.01242*, 2021.

[7] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International conference on machine learning*, pages 11525–11536. PMLR, 2021.

[8] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

[9] Fan Yang, Kai Wu, Shuyi Zhang, Guannan Jiang, Yong Liu, Feng Zheng, Wei Zhang, Chengjie Wang, and Long Zeng. Class-aware contrastive semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14421–14430, 2022.

[10] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[11] Jingyang Li, Jiachun Pan, Vincent YF Tan, Kim-Chuan Toh, and Pan Zhou. Towards understanding why fixmatch generalizes better than supervised learning. *arXiv preprint arXiv:2410.11206*, 2024.

[12] Qiyu Liao, Xin Yuan, Min Xu, and Dadong Wang. Sgia: Enhancing fine-grained visual classification with sequence generative image augmentation. *arXiv preprint arXiv:2412.06138*, 2024.

[13] Hongyang He, Hongyang Xie, Haochen You, and Victor Sanchez. Semi-vim: Bidirectional state space model for mitigating label imbalance in semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 765–774, 2025.

[14] Hongyang He, Hongyang Xie, Guodong Shen, Boyang Fu, Haochen You, and Victor Sanchez Silva. 4s-classifier: empowering conservation through semi-supervised learning for rare and endangered species, 2025.

[15] Wenjie Dang, Shuiwang Li, Qijun Zhao, and Fang Liu. Learning disentangled representation for fine-grained visual categorization. In *Image and Graphics: 11th International Conference, ICIG 2021, Haikou, China, August 6–8, 2021, Proceedings, Part I 11*, pages 327–339. Springer, 2021.

[16] Hongyang He, Xinyuan Song, Yangfan He, Zeyu Zhang, Yanshu Li, Haochen You, Lifan Sun, and Wenqiao Zhang. Trico: Triadic game-theoretic co-training for robust semi-supervised learning. *arXiv preprint arXiv:2509.21526*, 2025.

[17] Bowen Tian, Songning Lai, Lujundong Li, Zhihao Shuai, Runwei Guan, Tian Wu, and Yutao Yue. Pepl: Precision-enhanced pseudo-labeling for fine-grained image classification in semi-supervised learning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[18] Yue Duan, Zhen Zhao, Lei Qi, Luping Zhou, Lei Wang, and Yinghuan Shi. Roll with the punches: expansion and shrinkage of soft label selection for semi-supervised fine-grained learning, 2024.

[19] Manuel Lagunas, Brayan Impata, Victor Martinez, Virginia Fernandez, Christos Georgakis, Sofia Braun, and Felipe Bertrand. Transfer learning for fine-grained classification using semi-supervised learning and visual transformers. *arXiv preprint arXiv:2305.10018*, 2023.

[20] Andrew Kyle Lampinen, Stephanie CY Chan, and Katherine Hermann. Learned feature representations are biased by complexity, learning order, position, and more. *arXiv preprint arXiv:2405.05847*, 2024.

[21] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

[22] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[23] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[24] Junbo Zhang and Kaisheng Ma. Rethinking the augmentation module in contrastive learning: Learning hierarchical augmentation invariance with expanded views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16650–16659, 2022.

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.

[26] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[27] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset, 2011.

[28] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[29] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 595–604, 2015.

[30] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[31] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

A. Proof of Theorem 1

Proof. Fix $p \in \{1, \dots, P\}$. Regard the loss as a function of the p-th token only and write

$$L(x_p) = \mathcal{L}_{CE}(X_u, \hat{y}_u), \quad I_p = \left\| \frac{\partial L(x_p)}{\partial x_p} \right\|_2. \tag{16}$$

After masking, the p-th token becomes $\tilde{x}_p = \beta_p x_p$ and the corresponding loss is

$$\tilde{L}(\tilde{x}_p) = \mathcal{L}_{CE}(\tilde{X}_u, \hat{y}_u), \tag{17}$$

where $\tilde{X}_u$ is obtained from X_u by replacing x_p with $\tilde{x}_p$ and keeping all other tokens fixed.

Let $g(x_p) = \frac{\partial L(x_p)}{\partial x_p}$ and $H(x_p) = \frac{\partial^2 L(x_p)}{\partial x_p^2}$. By assumption, $\|H(x_p)\| \le \sup_X \left\| \frac{\partial^2 \mathcal{L}_{CE}(X_u, \hat{y}_u)}{\partial x_p^2} \right\|$ for all relevant X. Define

$$M = \sup_X \left\| \frac{\partial^2 \mathcal{L}_{CE}(X_u, \hat{y}_u)}{\partial x_p^2} \right\| \cdot \|x_p\|_2. \tag{18}$$

Consider the gradient at the masked token $\tilde{x}_p$. By the mean value theorem in $\mathbb{R}^d$,

$$g(\tilde{x}_p) - g(x_p) = \int_0^1 H\big(x_p + t(\tilde{x}_p - x_p)\big)\big(\tilde{x}_p - x_p\big)\, dt. \tag{19}$$

Since $\tilde{x}_p = \beta_p x_p$, we have $\tilde{x}_p - x_p = (\beta_p - 1)x_p$, and hence

$$\begin{aligned} \|g(\tilde{x}_p) - g(x_p)\|_2 &\le \int_0^1 \big\|H\big(x_p + t(\beta_p - 1)x_p\big)\big\| \cdot |\beta_p - 1| \cdot \|x_p\|_2 \, dt \\ &\le (1 - \beta_p)M, \end{aligned} \tag{20}$$

because $0 < \beta_p \le 1$ implies $|\beta_p - 1| = 1 - \beta_p$ and $\|H(\cdot)\|$ is uniformly bounded by the definition of M. Therefore,

$$\begin{aligned} \|g(\tilde{x}_p)\|_2 &\le \|g(x_p)\|_2 + \|g(\tilde{x}_p) - g(x_p)\|_2 \\ &\le I_p + (1 - \beta_p)M. \end{aligned} \tag{21}$$

Now, let us relate this to the gradient with respect to the original variable x_p after masking. Since $\tilde{x}_p = \beta_p x_p$, the chain rule gives

$$\begin{aligned} \frac{\partial \mathcal{L}_{CE}(\tilde{X}_u, \hat{y}_u)}{\partial x_p} &= \frac{\partial \tilde{L}(\tilde{x}_p)}{\partial \tilde{x}_p} \cdot \frac{\partial \tilde{x}_p}{\partial x_p} \\ &= g(\tilde{x}_p) \cdot \beta_p. \end{aligned} \tag{22}$$

Taking norms and using the previous bound,

$$\begin{aligned} \left\| \frac{\partial \mathcal{L}_{CE}(\tilde{X}_u, \hat{y}_u)}{\partial x_p} \right\|_2 &= \beta_p \|g(\tilde{x}_p)\|_2 \\ &\le \beta_p I_p + \beta_p(1 - \beta_p)M. \end{aligned} \tag{23}$$

Finally, since $0 < \beta_p \le 1$ implies $\beta_p(1 - \beta_p) \le 1 - \beta_p$, we obtain

$$\left\| \frac{\partial \mathcal{L}_{CE}(\tilde{X}_u, \hat{y}_u)}{\partial x_p} \right\|_2 \le \beta_p I_p + (1 - \beta_p)M, \tag{24}$$

which is exactly the claimed inequality. □

B. Proof of Theorem 2

Proof. Define $L_{\text{orig}}(X_u) = \text{CE}(F(X_u), \hat{y}_u)$ and $L_{\text{cons}}(X_u) = \text{CE}(F(\tilde{X}_u), \hat{y}_u)$, where $\tilde{X}_u = T(X_u)$. By the chain rule,

$$\nabla_{X_u} L_{\text{cons}}(X_u) = \frac{\partial L_{\text{cons}}}{\partial \tilde{X}_u} \cdot \frac{\partial \tilde{X}_u}{\partial X_u}. \tag{25}$$

Write $\Delta(X_u) = \tilde{X}_u - X_u$, and assume $\|\Delta(X_u)\|_F \leq \delta$. If T is differentiable, then $\frac{\partial \tilde{X}_u}{\partial X_u} = I + J_T(X_u)$, where J_T is the Jacobian of T. Using the composite structure $L_{\text{cons}}(X_u) = \text{CE}(F(T(X_u)), \hat{y}_u)$, the multivariate chain rule yields

$$\nabla_{X_u} L_{\text{cons}}(X_u) = J_T(X_u)^\top \nabla_{\tilde{X}_u} \text{CE}(F(\tilde{X}_u), \hat{y}_u). \tag{26}$$

Let $g(X) = \nabla_X \text{CE}(F(X_u), \hat{y}_u)$. The second-order Taylor expansion of $g(\tilde{X}_u)$ around X_u gives

$$g(\tilde{X}_u) = g(X_u) + \nabla g(\xi) \cdot \Delta(X_u), \tag{27}$$

where ξ lies on the segment between X_u and $\tilde{X}_u$, and the Hessian bound implies $\|\nabla g(\xi)\|_F \leq H_{\max}$. Hence

$$\nabla_{X_u} L_{\text{cons}}(X_u) = J_T(X_u)^\top \left[g(X_u) + \nabla g(\xi) \cdot \Delta(X_u) \right]. \tag{28}$$

Since $\nabla_{X_u} L_{\text{orig}}(X_u) = g(X_u)$, the difference becomes

$$\nabla_{X_u} L_{\text{cons}}(X_u) - \nabla_{X_u} L_{\text{orig}}(X_u) = \left(J_T(X_u)^\top - I \right) g(X_u) + J_T(X_u)^\top \nabla g(\xi) \cdot \Delta(X_u). \tag{29}$$

Taking Frobenius norms and applying the triangle inequality,

$$\left\| \nabla_{X_u} L_{\text{cons}} - \nabla_{X_u} L_{\text{orig}} \right\|_F \leq \left\| J_T(X_u)^\top - I \right\|_F \cdot \|g(X_u)\|_F + \left\| J_T(X_u)^\top \right\|_F \cdot \|\nabla g(\xi)\|_F \cdot \|\Delta(X_u)\|_F. \tag{30}$$

Assume T is a mild augmentation so that $\|J_T(X_u)^\top - I\|_F \leq \epsilon$ and $\|J_T(X_u)^\top\|_F \leq 1 + \epsilon$. Using $\|\Delta(X_u)\|_F \leq \delta$, $\|\nabla g(\xi)\|_F \leq H_{\max}$, and $\|g(X_u)\|_F = \|\nabla_{X_u} L_{\text{orig}}\|_F$, we obtain

$$\left\| \nabla_{X_u} L_{\text{cons}} - \nabla_{X_u} L_{\text{orig}} \right\|_F \leq \epsilon \left\| \nabla_{X_u} L_{\text{orig}} \right\|_F + (1 + \epsilon) H_{\max} \delta. \tag{31}$$

Let $\epsilon \leq \delta L$, where L is the Lipschitz constant of ∇F. Substituting and dropping higher-order δ^2 terms yields

$$\left\| \nabla_{X_u} L_{\text{cons}} - \nabla_{X_u} L_{\text{orig}} \right\|_F \leq \delta \left(L \left\| \nabla_{X_u} L_{\text{orig}} \right\|_F + H_{\max} \right). \tag{32}$$

By the triangle inequality,

$$\|\nabla_{X_u} L_{\text{cons}}\|_F \leq \left\| \nabla_{X_u} L_{\text{orig}} \right\|_F + \delta \left(L \left\| \nabla_{X_u} L_{\text{orig}} \right\|_F + H_{\max} \right). \tag{33}$$

Using the model Lipschitz property $\|\nabla_{X_u} L_{\text{orig}}\|_F \leq L \|X_u\|_F$ and absorbing constant factors into L and $H_{\max}$, we obtain the final bound

$$\|\nabla_{X_u} \mathcal{L}_{\text{token-cons}}\|_F \leq \left\| \nabla_{X_u} \mathcal{L}_{\text{pseudo}} \right\|_F + \delta \left(L + H_{\max} \cdot \|X_u\|_F \right). \tag{34}$$

This completes the proof. □