Pattern Recognition 183 (2027) 114994



Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr





TeD−Loc: Text distillation for weakly supervised object localization

Shakeeb Murtaza *, Soufiane Belharbi, Alexis Guichemerre, Marco Pedersoli, Eric Granger

Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA), Dept. of Systems Engineering, ETS Montreal, Canada

ARTICLE INFO

Keywords:
Deep learning
Weakly supervised object localization
Self-supervised learning
Vision language models
Knowledge distillation

ABSTRACT

Weakly supervised object localization (WSOL) models can predict both the object class and the spatial regions corresponding to the object, without requiring explicit bounding-box annotations. Given their reliance on classification objectives, traditional WSOL methods, like class activation mapping, tend to focus on the most discriminative object regions, often missing the full spatial extent. Although vision–language models like CLIP encode rich semantic priors, their global text and class-token embeddings are not explicitly aligned with local patch embeddings, limiting patch-level localization. Recent methods such as GenPrompt address this limitation, but at the cost of increased complexity, as they rely on conditional denoising and elaborate prompt-learning strategies. In this paper, we propose Text Distillation for Localization (TeD−Loc), which distills knowledge from CLIP text embeddings to patch embeddings through contrastive alignment, thereby enabling patch-level foreground/background localization. A localization-guided classification module is also introduced, which uses localization scores to aggregate foreground patch embeddings for joint classification and localization within a single model. In addition, a QR-based orthogonalization of class text embeddings is applied before distillation to improve discrimination for semantically similar classes. Extensive experiments[1] show that TeD−Loc improves Top−1 Loc by ∼5% on CUB and ILSVRC, and PxAP by ∼31% on histopathology benchmarks, while achieving more efficient inference than GenPrompt.

1. Introduction

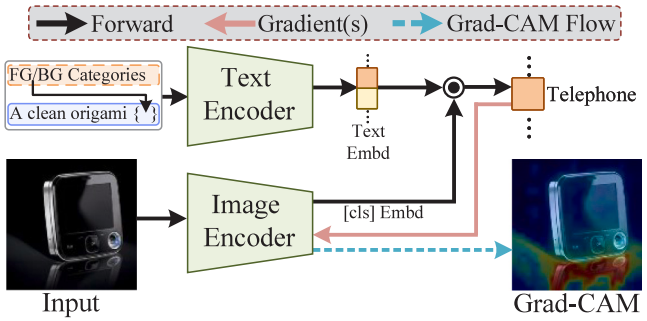
WSOL is a critical yet challenging task in computer vision, aiming to localize objects within images using a model trained using only image-class labels rather than instance-level annotations. The popular class activation mapping (CAM) method [1] leverages classification models for generating localization maps. However, they inherently focus on the most salient regions of an object and often fail to capture the full spatial extent [2]. This limitation arises because discriminative models are optimized to minimize mutual information between different instances of the same class. Various strategies have been proposed to mitigate this issue, including spatial regularization [3,4], adversarial erasing [5–7], and leveraging pseudo-labeling [2,8,9]. However, these approaches are constrained by the local receptive fields of convolutional neural networks (CNNs), which limit their ability to capture global dependencies essential for complete object localization.

Vision transformers (ViTs) [10] have recently shown potential in modeling long-range dependencies through self-attention. This addresses key limitations of CNNs for WSOL — the restricted local receptive fields that prevent capturing global spatial dependencies, and the partial activation issue, whereby only the most discriminative object regions are highlighted [2]. However, they lack the local inductive biases inherent to CNNs, often resulting in weaker local feature representations. Vision–language models, particularly contrastive language–image pre-training (CLIP), have emerged as a promising direction by aligning textual and visual features, which can be leveraged for localization tasks using class-level labels [11]. Yet, an architectural limitation of transformer-based models is the disconnection between the text embeddings and the local patch embeddings. Since the class token aggregates global image semantics rather than spatially grounded representations, it remains poorly aligned with the local features required for localization, making it challenging to extract accurate localization maps. Moreover, the predominant approaches for extracting localization maps from CLIP, such as gradient-based and attention manipulation methods, rely heavily on ground truth (GT) class information. This dependency leads to performance degradation when predicted classes derived from an external model are employed because of feature misalignment and class confusion, as illustrated by CLIP's frequent conflation of similar classes like "airplane" and "aircraft" [12]. Consequently, employing a CLIP model without fine-tuning introduces substantial errors in downstream tasks, prompting the need for strategies that can learn precise localization cues while minimizing reliance on explicit class labels.

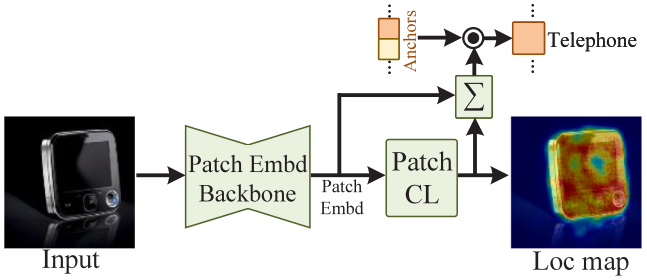
* Corresponding author.
E-mail address: shakeeb.murtaza.1@ens.etsmtl.ca (S. Murtaza).

[1] Our code is available at: github.com/shakeebmurtaza/TeDLOC.

https://doi.org/10.1016/j.patcog.2026.114994
Received 16 November 2025; Received in revised form 14 July 2026; Accepted 18 September 2026
Available online 23 September 2026
0031-3203/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).



(a) CLIP-ES localization via standard Grad-CAM [11] at test time.



(b) TeD-Loc localizes an object without class labels at test time.

Fig. 1. A comparison of our TeD-Loc versus CLIP-ES [11] methods for extracting localization maps from CLIP at test time. (A) CLIP-ES utilizes Grad-CAM to extract localization maps from CLIP, requiring GT class labels during inference. (B) In contrast, our TeD-Loc model distills knowledge from CLIP text embeddings into the visual encoder during training, allowing it to produce both classification scores and localization maps without requiring class labels during inference.

Recently, GenPrompt [13] attempted to address these challenges by framing WSOL as a conditional denoising process, leveraging CLIP embeddings to capture discriminative regions. While GenPrompt improves localization by using CLIP embeddings, it still relies on external classifiers or GT class labels during inference, adding to their complexity and limiting their applicability in real-world scenarios. Moreover, despite robust map generation capabilities, a fundamental limitation remains the inability of CLIP-based methods to localize objects within an image without prior class information. This constraint poses a significant challenge for downstream tasks, as they require computing class labels beforehand (Fig. 1(a)). Given these challenges, we seek to effectively harness vision–language models to learn precise localization cues for WSOL while mitigating misclassification and reducing reliance on GT labels during inference.

To address these challenges, we propose Text Distillation for Localization (TeD-Loc). CLIP's text embeddings encode semantic information that aligns strongly with visual concepts. However, CLIP cannot align text embeddings with local patch-level representations, limiting its ability to localize objects. TeD-Loc addresses this gap by distilling knowledge from CLIP's text embeddings, which serve as a powerful link between visual and textual modalities.

Localization information is learned by transferring knowledge from text embeddings to the localization module (Fig. 1(b)). Using contrastive learning within a teacher–student framework, the patch-level visual representations of our model are aligned with text embeddings. This alignment is guided by pseudo-labels that can be extracted using an off-the-shelf method based on CAMs. Learning from text embeddings allows TeD-Loc to achieve state-of-the-art performance using one model that is selected using the best localization performance without requiring separate classifiers trained and selected using the validation set over classifier scores. TeD-Loc introduces a new paradigm where classification is achieved through localization, thereby eliminating the need for model selection over classifier scores. By training the localization module to distinguish between foreground (FG) and background (BG) regions based on their similarity to text embeddings, TeD-Loc enables the model to classify and localize objects simultaneously. Furthermore, to address the limitations of CLIP's frequent conflation of semantically similar classes, we propose a method to orthogonalize text embeddings before distillation. By default, text embeddings in CLIP may not be sufficiently discriminative between similar classes because of their proximity in the embedding space. To mitigate this issue, we decompose the embeddings using QR decomposition [14] and utilize the resulting orthogonal basis vectors for alignment.

Our TeD-Loc method employs a transformer-based architecture that decomposes an image into a set of patches, generating upsampled patch embeddings through our model backbone to produce a fine-grained localization map. Each patch embedding is individually classified to estimate its likelihood of representing a FG or BG region. These classification scores are then gathered to produce a localization map, highlighting regions of interest within the image. The global classification score for the entire image is a weighted average of the patch embeddings. Our approach is inspired by the multiple instance learning (MIL) framework [15], where each image consists of a "bag" containing multiple "instances" (image patches) with only bag-level labels available during training. Leveraging MIL, object localization and classification are performed simultaneously by assigning higher weights to discriminant patches, without relying on external classifiers or prior class information. This aligns well with WSOL, enabling our model to independently produce accurate localization maps and classification scores.

Our main contributions are summarized as follows.

(1) The TeD-Loc method is introduced that distills knowledge from CLIP text embeddings into a patch embedding using contrastive learning. It allows for patch-level foreground/background localization without requiring class labels or external classifiers at inference time.

(2) A classification module is introduced that leverages localization scores to compute the expected embeddings of FG regions. Using a weighted average of patch embeddings, where the weights are derived from the FG localization map. This ensures that FG embeddings align with the correct class embeddings. This eliminates the need for an external classifier and allows our model to classify and localize simultaneously.

(3) To mitigate the tendency of CLIP to confuse semantically similar classes, a QR-based orthogonalization of class text embeddings is proposed before distillation, improving discriminability for both localization and classification.

(4) An extensive set of experiments on natural (CUB and ILSVRC) and histology (GlaS and CAMELYON17) image datasets indicate that TeD-Loc can outperform state-of-the-art WSOL methods on several challenging benchmarks.

2. Related work

(a) Weakly supervised object localization. WSOL is a challenging task that seeks to localize objects using only image-level supervision. The foundational work in WSOL [1] proposes to harvest CAMs from pre-trained CNNs, leveraging global average pooling (GAP) to guide the network's attention toward specific regions in an image. Despite its impact, CAM and related CNN-based approaches are constrained to highlight discriminative regions, often neglecting complete object extents. This limitation has led to the development of different WSOL methods that can look beyond discriminative regions. Furthermore, [16] proposes to utilize inter-class and intra-class gradients to improve CAMs.

Erasing-based methods aim to mitigate CAM partial activation by selectively obscuring parts of an image to encourage broader localization. HaS [17] and CutMix [18] employ random erasure, which forces the network to explore different object parts beyond discriminative regions. Building on this, adversarial erasing methods like ACoL [5] and ADL [6] use dual classifiers to identify and erase dominant regions, uncovering complementary object regions. Techniques like SPG [8] go further, integrating pixel-wise correlation constraints to maintain context and consistency across object regions.

Other works target the inherent challenge in CNNs to capture only local semantic features due to limited receptive fields. Consequently, newer methods leverage structural cues and integrate BG suppression techniques. SPA [19] enhances structural consistency, while PSOL [20] introduces a two-stage WSOL approach that decouples classification from localization tasks, providing robust pseudo-annotations for regression without class constraints. Methods such as BAS [3] reinforce this separation by suppressing BG regions and emphasizing FG areas critical to localization.

To overcome the inherent CAM limitations in capturing long-range dependencies, transformer-based approaches for WSOL are gaining traction. Transformers, known for their self-attention mechanism, enable networks to capture both local and global feature dependencies. More recently, CDTR [21] proposes embedding learnable category biases into the self-attention maps of a vision transformer (ViT) to learn category-aware representations for accurate localization. In WSOL, TS-CAM [10] leverages token-patch fusion with semantic maps to improve spatial coherence. By exploiting the long-range capability of transformers, these methods significantly broaden the object localization scope and address the core issues of CNN-based localization methods. Extending WSOL to video, [22] proposes exploiting color cues across frames to improve localization maps. Beyond closed-set settings, [23] introduces open-world WSOL for unseen categories via contrastive representation co-learning.

(b) Contrastive language–image pre-training. CLIP [24] is a foundational model designed to align visual and language representations, trained on 400 million image-text pairs collected from web data. By learning from paired data, CLIP produces a robust model capable of zero-shot adaptation to diverse tasks by computing similarities between images and textual descriptions. In WSOL, GenPrompt [13] leverages CLIP to identify discriminative regions and employs VQGAN for embedding generation within a denoising process to localize objects; however, its computational complexity hinders real-time applicability.

Moreover, CLIP has been widely adapted for weakly supervised semantic segmentation (WSSS), prompting multiple approaches to leverage its capabilities for generating class activation maps (CAMs) without extensive annotation. For instance, CLIMS [25] uses CLIP to enhance the completeness of object regions within CAMs while suppressing BG regions. CLIP-ES [11] employs GradCAM to extract CAMs directly from CLIP, demonstrating that activations can be generated without fine-tuning. SCLIP [26] and NACLIP [27] propose modifying the last attention block to produce score maps for each patch embedding, enabling segmentation map generation without requiring a backward pass. Beyond attention manipulation, WeakCLIP [28] introduces a text-to-pixel matching paradigm for WSSS. It exploits CLIP knowledge to refine CAMs. Similarly, [29] introduces learnable prompts in CLIP and uses pseudo-labels from SAM [30] for fine-tuning for WSSS.

While CLIP-based methods yield competitive maps, these methods rely on hand-crafted textual templates and predefined class representations, such as prompts like “a photo of [CLS]”. This requires prior knowledge of the specific class name before producing a localization map of each image. This reliance restricts the model’s adaptability across different computer vision tasks. It also leads to a substantial decline in performance when using predicted classes due to feature misalignment and class confusion. Furthermore, the disconnection between CLIP’s text embeddings and its local patch embeddings limits the model’s ability to capture spatially grounded representations. These limitations underscore a critical challenge: the need for WSOL methods that can learn precise localization cues without reliance on explicit class labels or predefined textual templates. To address this challenge, we propose a novel method that distills knowledge from text–image representations of CLIP to guide the localization network. Additionally, we mitigate the tendency of CLIP to conflate semantically similar classes by orthogonalizing the text embeddings before alignment, reducing semantic overlap and improving discriminability. This enables the model to classify the image by computing the similarity between visual embeddings and class anchors as visual embeddings are pushed toward class-text embeddings.

3. The proposed TeD-Loc method

Let us consider a training set $\mathbb{D} = \{(x_i, y_i)\}_{i=1}^{N}$ of N images, where each image $x_i \in \mathbb{R}^{H\times W\times 3}$ is associated with an image-level label $y_i \in \{1, \dots, K\}$, representing one of K object classes, with no bounding box (bbox) supervision. WSOL methods seek to train a model for object localization and classification using only image-level labels. In this paper, we leverage text embeddings from a pre-trained vision–language model, specifically CLIP [27].

Our model (see Fig. 2) consists of a patch embedding backbone network and a compact head for localization and classification tasks. The backbone network is comprised of the (i) *Encoder* E, a ViT (ViT-EVA-L) [31] that decomposes images into patches and produces patch-level embeddings. It is pre-trained for classification and frozen during training. (ii) The *Decoder* D upscales these patch embeddings to a high spatial resolution where the pth output patch is denoted as $z_p \in \mathbb{R}^d$. The CLIP text encoder is denoted as E_T. Moreover, we introduce a binary patch classifier $g(z_p)$ that predicts the FG/BG for each patch. Its response across all patches forms a localization map $M \in [0,1]^{H\times W}$ that contains the FG object associated with the image class. This map provides the localization generated by our method. Furthermore, we define the classification function $f : \mathbb{R}^d \to [0,1]^K$, which is parametrized with *frozen* class weight vectors $t_k \in \mathbb{R}^d$ for a class k for $k \in \{1, \dots, K\}$. Given an embedding vector $v \in \mathbb{R}^d$, the function f returns a probability distribution over the K classes by applying a softmax to the dot-product similarities between v and all class anchors, $f(v) = \mathrm{softmax}(\langle v, t_1\rangle, \dots, \langle v, t_K\rangle)$. Therefore, its kth entry $f_k(v)$ is the predicted probability of class k. In this work, we consider t_k as the CLIP text embedding of the class k, while v could be a patch embedding z_p, or the global image embedding h of our method.

3.1. Key components

The rest of this subsection introduces two important components of TeD-Loc: the generation of pseudo-labels of patches and the preprocessing of text embeddings.

Patch-level pseudo-label generation. To train our model, we propose to leverage patch-level pseudo-annotations corresponding to FG and BG patches. However, since such annotations are not available in the WSOL setting, we consider an off-the-shelf pre-trained classification model with a CAM method. Such models can yield a discriminative map to localize a target class, which is adequate during training. Generally, any CAM-based model can be used [32–34] to generate localization cues that can be leveraged for training. During the training of our model, we randomly sample a few FG/BG patch locations for each training image at each training step to avoid overfitting. This pseudo-labeling technique has been effective in guiding WSOL learning [2,35]. In our experiments, we sample the same number of FG/BG locations to maintain a balanced ratio between the two classes. ω denotes the set containing the sampled patches for both FG/BG, while ω^+ contains only

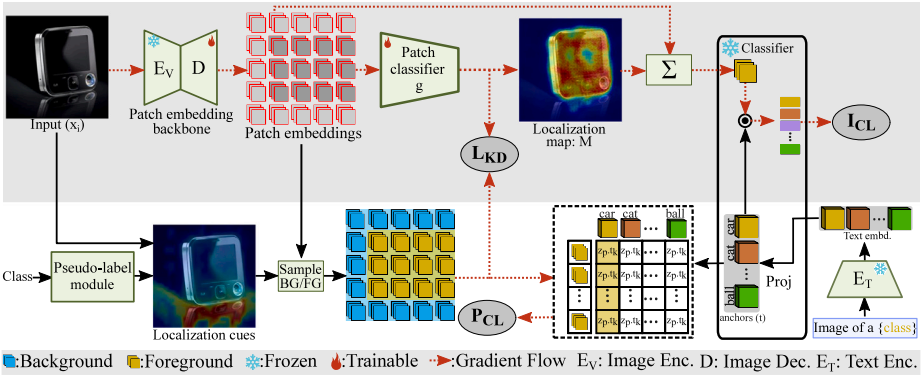


Fig. 2. An overview of the TeD-Loc method for distilling FG text embeddings into the patch embedding backbone. First, pseudo-labels are extracted to guide the identification of FG and BG patches. By leveraging these FG/BG regions, the model minimizes the similarity of E_V with the relevant text embedding for FG classes while maximizing dissimilarity with embeddings of other classes. Using a binary FG/BG classifier, TeD-Loc generates localization maps by classifying patches as FG or BG and producing class probabilities for image classification. This joint task enables the model to produce both accurate localization and classification outputs without explicit bounding box supervision.

the selected FG patches. $y'_p \in \{0, 1\}$ is used as the patch pseudo-label where 0 is BG and 1 is FG.

Sampling FG/BG regions for pseudo-label generation. To train our model, we employ pseudo-labels for FG and BG regions following recent methods [2,9,35]. We obtain these pseudo-labels by utilizing CAM $C \in \mathbb{R}^{H \times W}$ extracted from a pre-trained classifier model. These CAMs indicate regions of the image highlighting an object belonging to the FG class, which can be used to guide the sampling of FG and BG regions. We first apply Otsu's thresholding method [36] to the CAM C to separate FG and BG regions. This method automatically determines a threshold, effectively separating high-activation regions (FG) from low-activation ones (BG). We denote the set of pixel locations in the image domain as Ω.

For FG sampling, we focus on regions with high activation values in C. Specifically, we select the top n^+ pixels with the highest activation values inside the image, forming the set of potential FG locations $\omega_+ \subset \Omega$. We then randomly sample a subset of these locations to be used as FG samples during training. For BG sampling, we consider regions with low activation values in C. We select the bottom n^- pixels with the lowest activation values, excluding any pixels that are within the FG regions. This forms the set of potential BG locations $\omega_- \subset \Omega$. We randomly sample from these locations to obtain BG samples for training. We note that $\omega = \omega_+ \cup \omega_-$ is the set of all sampled pixels in both FG and BG in one sampling step.

To ensure that our model generalizes well and avoids overfitting, we perform this sampling process at every training step for each image. This dynamic sampling allows the model to explore different regions of the image during training, promoting robustness and consistency in learning.

The sampled FG and BG pixel locations are used to create a partial pseudo-label mask $y'_p \in \{0, 1\}$, where $y'_p = 1$ for FG pixels, $y'_p = 0$ for BG pixels, and locations with unknown labels are left undefined. The set of sampled locations p is defined by $p \in \omega$.

Text embeddings orthogonalization. CLIP text embeddings of classes are used to distill localization knowledge as they provide a powerful link between global visual and textual representations. However, text embeddings can sometimes conflate similar classes due to semantic overlap (e.g., "airplane" and "aircraft") [12]. This overlap between classes limits the benefit of those embeddings, especially when used in discriminative scenarios. To mitigate this issue, we propose to pre-process the class-text embeddings before using them in our method. We consider a transformation that projects all the text embeddings into a space where the distance between each pair of embeddings is maximum. In this work, we use orthogonal projection, in particular, QR orthogonalization [14] and conserve the basis of the projection. In the rest of this paper, the orthogonalized version of text embeddings for class k is referred to as t_k. These new embeddings are kept frozen and play the role of class anchors that are carried in our model, allowing us to discard the text encoder. Fig. 3 illustrates the issue of class overlap and the impact of orthogonalization over text embeddings.

3.2. TeD-Loc training method

Text embedding distillation to local patches. Our objective is to distill text-class embeddings into patch embeddings and to discard the text encoder. Since text and global image embeddings in CLIP are not directly tied to patch embeddings, text embeddings cannot directly be used to localize objects. Patch embeddings are not necessarily correlated with text embeddings [27]. In this work, we propose to create a direct link between text embeddings and patch embeddings for semantic localization. To achieve this, we propose to use knowledge distillation, where we transfer the text embedding of the image class to the patch embeddings, allowing us to perform object localization. In particular, a contrastive learning loss [24] is employed to ensure that FG patch embeddings are similar to the text embedding of the image class while being dissimilar from embeddings of other classes. To this end, only the embeddings z_p of FG patches, with $p \in \omega^+$, are used for this loss. It can be defined through a standard cross-entropy as follows [24],

$$\mathbf{L}_{\text{KD}} = \sum_{p \in \omega^+} CE(y, f(z_p)). \tag{1}$$

where $t_y \in \mathbb{R}^d$ denotes the orthogonalized text embedding of the GT class y, that is, the class anchor for the image. Concretely, for each foreground patch embedding z_p, $f(z_p)$ is the class distribution obtained by a softmax over the dot-product similarities $\langle z_p, t_k \rangle$ between z_p and *all* K class text anchors. The cross-entropy in Eq. (1) therefore trains z_p to maximize its dot-product similarity to the ground-truth anchor t_y while simultaneously minimizing similarity to every other anchor t_k, $k \neq y$; the correct class acts as the single positive and the remaining $K-1$ class embeddings act as negatives, which is precisely the one-vs-all contrastive objective used in CLIP [24]. Furthermore, Eq. (1) is computed only on the selected FG patches ω^+, as BG patches lack corresponding class embeddings, rendering them unsuitable for this contrastive loss. Since BG is not considered, this can lead to poor localization as the BG region is present in most images. To mitigate this issue, we introduce a patch binary FG/BG classifier g, which repels BG patch embeddings from FG text embeddings. It is trained using the

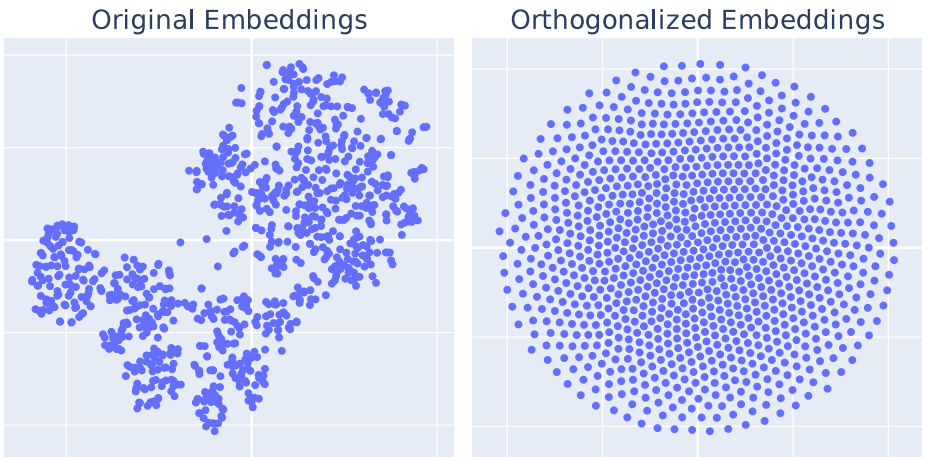


Fig. 3. t-SNE visualizations of CLIP text embeddings for ILSVRC [37] classes before and after orthogonalization. **(Left)** Prior to orthogonalization, embeddings of semantically similar classes (e.g., "airplane" and "aircraft") cluster closely together, leading to potential confusion. **(Right)** After orthogonalization (QR decomposition), the embeddings are more uniformly distributed and orthogonal, reducing overlap.

embeddings z_p of both FG and BG patches and their pseudo-labels y'_p via standard cross-entropy loss,

$$\mathbf{P}_{\text{CL}} = \sum_{p \in \omega} CE(y'_p, g(\mathbf{z}_p)). \tag{2}$$

Minimizing this loss allows for both FG and BG regions to be present in the image, which helps avoid imbalanced localization.

Global image embeddings from local patch embeddings for classification. So far, our method can only perform localization. To further allow it to perform image classification, our aligned FG patch embeddings are leveraged to construct a global image embedding that describes the object in the image. This creates a reversed link from local patch representations to global image representation, allowing us to learn to classify the image. This aligns perfectly with our distillation approach from class text embedding to patch embeddings described previously, where we ensure that FG patch embeddings are correlated with the text embedding of the image class. Therefore, we leverage this property to construct a global image embedding h using all the patch embeddings and the patch classifier g as follows,

$$h = \sum_p a_p z_p \quad \text{, where } a_p = g(z_p) / \sum_j g(z_j). \tag{3}$$

Eq. (3) performs a weighted average of the embeddings for all the patches by giving more importance to patches that are classified as FG since their $g(z_p)$ will be close to 1. In addition, BG patches are discarded since $g(z_p)$ is close to 0. This effectively performs a differentiable selection of FG patches, allowing for training with gradient-based methods. Most importantly, the final aggregated embedding $h \in \mathbb{R}^d$ is expected to resemble the text embedding of the image class t_k. To furthermore ensure this, this embedding is trained to be as close as possible to t_k using standard cross-entropy as follows,

$$\mathbf{I}_{\text{CL}} = CE(y, f(h)). \tag{4}$$

Overall training loss. Our overall training loss contains the three terms discussed previously: knowledge-distillation loss ($\mathbf{L}_{\text{KD}}$), patch classifier loss ($\mathbf{P}_{\text{CL}}$) and global image classification loss ($\mathbf{I}_{\text{CL}}$) as follows,

$$\mathbf{L} = \lambda_1 \mathbf{L}_{\text{KD}} + \lambda_2 \mathbf{P}_{\text{CL}} + \lambda_3 \mathbf{I}_{\text{CL}} \,, \tag{5}$$

where λ_1, λ_2 and λ_3 are weighting factors that balance the contribution of each term. Stochastic Gradient Descent (SGD) is used for optimizing the parameters of our model (parameters of the decoder D_I and patch classifier g). By jointly optimizing these loss functions, our model learns to produce discriminative and well-aligned visual representations, enabling simultaneous classification and localization.

4. Results and discussion

4.1. Experimental methodology

Dataset. `TeD-Loc` is evaluated on challenging natural and histology image benchmarks to assess its performance across standard WSOL scenarios.

Natural images: Two common challenging datasets were used for our WSOL experiments. (i) Caltech-UCSD birds-200-2011 (`CUB`) [38] consists of 11,788 images spanning 200 bird species. The dataset is partitioned into 5994 training images and 5794 testing images. For validation, an independent set of 1000 images (five per class) collected by [32] is utilized. (ii) ImageNet large-scale visual recognition challenge (`ILSVRC`) [37] includes approximately 1.2 million training images and 10,000 validation images across 1000 classes. We use the original validation split as our test set due to its sufficient size for robust evaluation. For validation purposes, `ILSVRC-V2`, collected by [39] and annotated by [32], is employed to mitigate biases toward the test set. For a fair comparison, we strictly adhere to the commonly used WSOL protocol proposed in [32] for both datasets.

Histology images: We evaluate `TeD-Loc` on two histopathology benchmarks covering distinct clinical settings: `GlaS` for colon cancer and `CAMELYON17` for breast cancer. `GlaS` provides colon histology images with gland-level annotations [34] and contains 67 training images, 18 validation images for classification, 6 validation images for localization, and 80 test images. For breast cancer, we use a recent WSOL benchmark derived from `CAMELYON17` [40], in which the original whole-slide dataset is reformulated into a patch-level setting suitable for weakly supervised classification and localization. Following [40], for each center and each class, 8/2/2 whole-slide images are selected for training, validation, and testing, respectively. Tumor and normal patches are then extracted using the provided tumor masks, while only informative tissue regions are retained to avoid trivial patches, as defined in [34,40]. The numbers of training, test, and validation images used for evaluation for each `CAMELYON17` center are summarized in Table 1 as introduced by [40].

Evaluation measures. Following earlier work by Choe et al. [32], we employ three localization measures alongside a classification measure to evaluate the proposed and baseline methods. The localization measures are as follows. (1) `MaxBoxAcc` (referred to in previous work as `CorLoc` [41] or `GT-Known` [17]), which quantifies the proportion of images for which the predicted bbox achieves an Intersection over Union (`IoU`) threshold of $\sigma = 50\%$ with the ground-truth bbox, independent of classification accuracy (`CL`); (2) Top-1 localization accuracy

Table 1
Dataset splits used in [40] for CAMELYON17.

	GlaS	C17-0	C17-1	C17-2	C17-3	C17-4
Train	67	634	1066	498	816	940
Val (CL)	18	144	38	110	102	146
Val (PxAP)	6	10	10	10	10	10
Test	80	262	172	448	376	298

(Top-1 Loc), measuring the proportion of images where the model's top predicted class is correct and the bbox IoU with ground truth exceeds $\sigma = 50\%$; and (3) Top-5 localization accuracy (Top-5 Loc), defined as the proportion of images for which the true class label is within the model's top-five predictions and the bbox meets an IoU of $\sigma = 50\%$. For histology, to evaluate localization performance, we use PxAP, a metric introduced in [32] that is based on pixel-wise precision and recall and has been widely used in histology benchmarks [34,40,42].

Implementation details. We closely followed the experimental setup of Choe et al. [32], dataset splits, evaluation of localization maps across multiple thresholds, and training epochs. Specifically, 50 epochs for the CUB dataset and 10 epochs for ILSVRC. Furthermore, our model is trained with a batch size of 32 and 16 for ILSVRC and CUB, respectively. In Eq. (5), the hyper-parameters λ_1, λ_2 and λ_3 used in the total training loss (Eq. (5)) terms that are optimized over the values $(0, 1]$. Optimization of our model was performed using SGD, with a learning rate from 1e-6 up to 0.01. We also fine-tuned the weight decay and momentum. In our experiments, we use EVA-CLIP pre-trained on natural images together with CLIP-ES as the CAM module. Localization maps were evaluated at a resolution of 256×256.

For histology, the compared WSOL baseline methods are trained for 1000 epochs on GlaS and 500 epochs on CAMELYON17. We use a weight decay of 0.0001. During training, images are first resized to 256×256 and then randomly cropped to 224×224, with the learning rate selected from {0.0001, 0.001} and the learning-rate decay factor fixed to 0.9. For TeD-Loc, we follow the same training setup as in the natural-image setting. Specifically, images are first resized to 512×512 and then randomly cropped to 448×448. Due to the size of the model, training is performed for 500 epochs on GlaS and 200 epochs on CAMELYON17. Following prior work [2,42], we perform a hyper-parameter search over the loss weights λ_1, λ_2 and λ_3 in {1, 2}, and over the learning rate in {0.0001, 0.001}. We use the pretrained CONCH vision–language model [43], including both its visual and text encoders. To ensure direct comparability, we adopt the same experimental protocol, evaluation metrics, and baseline methods as PixelCAM [44]. Moreover, applying TeD-Loc to histology requires a slight modification of its original training strategy. In its initial form, TeD-Loc aligns foreground pixel embeddings with the text embedding of the corresponding class. For CAMELYON17, this assumption does not hold for the normal class, since normal patches do not contain foreground regions. We therefore modify the alignment strategy by associating background embeddings with the text embedding of the normal class. To avoid an excessive number of background embeddings relative to foreground ones, we follow the strategy used in PixelCAM [44] and retain a proportion ρ of background embeddings, with ρ fixed to 0.5 in all experiments. For GlaS, we use the original TeD-Loc setting, since both normal and cancer images contain foreground/background tissue for alignment.

Baseline methods. For natural-image benchmarks, we compare TeD-Loc with recent state-of-the-art WSOL methods, including TS-CAM [10], SCM [45], LCTR [46], C^2AM [47], PSOL [20], DiPS [35], CATR [48], DA-WSOL [49], BAS [3], and GenPrompt [13]. Additionally, CLIP-ES [11], which utilizes Grad-CAM to extract localization maps from CLIP, was employed in a zero-shot setting. Specifically, we considered two variants of this method: CLIP-ES (GT-Known) and CLIP-ES (Pred). The CLIP-ES (GT-Known) variant requires ground-truth class labels during inference to generate localization maps. While this provides an upper bound on performance, it relies on privileged information unavailable in practical WSOL scenarios, thereby limiting its applicability. In contrast, CLIP-ES (Pred) depends solely on predicted class labels, aligning with the standard weakly supervised setting and offering a fair basis for comparison. This comprehensive evaluation enables us to demonstrate the robustness of our method across diverse settings.

For histology benchmarks, we compare against the histology WSOL baselines used in PixelCAM [44], including DeepMIL [50], GradCAM++ [51], LayerCAM [52], SAT [53], and PixelCAM [44]. As TeD-Loc relies on a CAM module for supervision, we use PixelCAM [44] by default to generate the pseudo-label CAMs. We further conduct an ablation study to evaluate the impact of the CAM module by replacing PixelCAM [44] with alternative WSOL methods, including DeepMIL [50], GradCAM++ [51], and LayerCAM [52].

4.2. Comparison with state-of-the-art methods

4.2.1. Evaluation on natural-image datasets

Quantitative results. The results in Table 2 show that TeD-Loc consistently improves localization performance over previous WSOL methods on both ILSVRC and CUB. On ILSVRC, our best variant (TeD-Loc ⟨patch,anchor⟩) attains a MaxBoxAcc of 77.1%, a Top-1 Loc of 71.8%, and a Top-5 Loc of 76.7%. This corresponds to absolute gains of 2.1, 6.6, and 3.3 percentage points, respectively, compared to the previous best reported results from GenPrompt [13]. On CUB, TeD-Loc reaches a MaxBoxAcc of 98.7%, a Top-1 Loc of 92.0%, and a Top-5 Loc of 97.5%, improving on GenPrompt by 0.7, 5.0, and 1.4 percentage points, respectively. The significant improvements on both datasets highlight the capability of TeD-Loc in handling diverse and challenging scenarios.

In terms of classification, TeD-Loc also yields a substantial improvement over the distillation backbone. On ILSVRC, our method attains a Top-1 CL accuracy of 89.9%, whereas the underlying CLIP model reaches 67.3%. On CUB, TeD-Loc achieves 93.0% Top-1 CL accuracy compared to 46.4% for CLIP. Thus, a single network provides improved performance in both localization and classification, supporting the hypothesis that transferring text embeddings to patch representations yields a shared representation that is discriminative for both tasks.

To show the efficacy of our proposed method, we compare its model complexity and inference speed with GenPrompt [13] (Table 3). Although GenPrompt yields comparable localization performance, it introduces significant computational overhead. More specifically, GenPrompt employs a generative framework involving diffusion models and CLIP embeddings to generate Loc maps, resulting in a considerable increase in complexity due to the multiple-module pipeline and iterative nature of the diffusion process. Specifically, it comprises an EfficientNet-based classifier for label prediction (66.35 million parameters), a variational autoencoder (VAE) for latent embeddings (83.65 million parameters), and CLIP E_T for discriminative and representation embeddings, having 123.83 million parameters. The diffusion model, utilizing a U-Net architecture, adds 859.5 million parameters and iterates across 100-time steps, drastically increasing computation time. Our proposed method thus emerges as a significantly more efficient alternative, achieving robust performance without the exorbitant computational cost characteristic of GenPrompt. In contrast, TeD-Loc is a compact model (569.67M parameters) and obviates the need for external classifiers or diffusion sampling. Furthermore, for inference time, we utilized an idle machine equipped with an NVIDIA A100 GPU. We first conducted 20 warm-up epochs with a batch size of 1, followed by 1000 inference steps, and calculated the average inference time across these steps.

Table 2
MaxBoxAcc, Top-1 Loc, and Top-5 Loc performance of TeD-Loc against state-of-the-art methods on the ILSVRC and CUB datasets. The first row corresponds to Grad-CAM for CLIP [11], where class labels are required to produce text embeddings for extracting localization maps. Notably, TeD-Loc outperforms existing methods without relying on text-encoder. The row TeD-Loc * (⟨patch, anchor⟩) reports a GT-known analysis variant, where localization maps are computed from patch–class similarity scores via the dot product $\langle z_p, t_y \rangle$. Here, z_p denotes the patch embedding at location $p \in \Omega$, and t_y is the text anchor associated with the ground-truth class y. This variant is used to assess the quality of the text-to-patch alignment.

Method	ILSVRC			CUB		
	MaxBoxAcc	Top-1 Loc	Top-5 Loc	MaxBoxAcc	Top-1 Loc	Top-5 Loc
CLIP-ES (Pred) [11] *(cvpr,2023)*	71.2	–	–	91.6	–	–
TS-CAM [10] *(iccv,2021)*	67.7	53.4	64.3	71.3	83.8	87.7
SCM [45] *(eccv,2022)*	68.8	56.1	66.4	76.4	91.6	96.6
LCTR [46] *(aaai,2022)*	68.7	56.1	65.8	92.4	79.2	89.9
PSOL [20] *(cvpr,2020)*	66.3	58.0	65.0	91.8	80.9	90.0
C^2AM [47] *(cvpr,2022)*	68.5	59.6	67.1	92.9	81.8	91.1
GenPrompt [13] *(iccv,2023)*	75.0	65.2	73.4	98.0	87.0	96.1
CATR [48] *(iccv,2023)*	69.2	56.9	66.6	94.9	79.6	92.0
DA-WSOL [49] *(pami,2024)*	71.8	55.3	–	88.4	71.1	–
BAS [3] *(ijcv,2022)*	72.0	58.5	69.0	94.6	72.0	88.1
SeCM [54] *(pr,2025)*	72.9	61.7	70.9	98.0	81.9	94.2
TeD-Loc (ours)	**75.6**	**70.0**	**75.1**	**98.7**	**91.7**	**97.6**
TeD-Loc (<patch,anchor>)	**77.1**	–	–	**98.7**	–	–

Table 3
Computational complexity of our proposed TeD-Loc against GenPrompt.

Methods	Complexity analysis	
	# Parameters	Inference time
GenPrompt [13]	1133.35M	272 ms
TeD-Loc (ours)	569.67M	121 ms

Qualitative results. Fig. 4 presents a visual comparison of our method against state-of-the-art WSOL approaches, specifically GenPrompt [13] and TS-CAM [10], on the ILSVRC dataset. While these methods yield competitive quantitative performance, they often struggle with accurately localizing complex objects. They tend to highlight irrelevant parts or even entirely different objects, especially in intricate scenes. GenPrompt, for instance, relies on CLIP's discriminative and representative embeddings during inference, which necessitates class labels at test time using the external classifier. While this approach aims to adaptively focus on the object of interest, it can erroneously localize incorrect objects when the classifier predicts the wrong class. This dependency on class information during inference increases the chances of mistakes, especially in critical applications where precise localization without prior class knowledge is required. In contrast, our method consistently achieves high localization accuracy by effectively capturing both discriminative and non-discriminative regions of the target object. These localization maps can produce bboxes that encompass the entire object, enhancing both localization performance and interpretability. Unlike other methods that produce low-activation regions resulting in bboxes over areas without meaningful content, our approach ensures that activations correspond closely with the actual visual appearance of the object. Furthermore, Fig. 4 shows different failure patterns for the baseline methods. For instance, TS-CAM highlight the most salient co-occurring object rather than the target (e.g., the child rather than the toilet seat), reflecting its partial-activation bias. Also, GenPrompt fails when its external classifier mispredicts the class, since an incorrect label (from the external classifier) propagates a localization. TeD-Loc able to deal with these fail modes as the distilled patch embeddings carry semantic information, producing sharp activation maps across different scenes.

4.2.2. Evaluation on histology datasets

To assess the generalization of TeD-Loc beyond natural-image benchmarks, we evaluate it on two challenging histology datasets: GlaS for colon cancer and CAMELYON17 for breast cancer, which includes data from five centers. Unlike natural images, histology images present distinct challenges, as regions of interest often resemble the surrounding background and substantial variability may exist within the same class [42,44]. Consequently, WSOL methods tend to perform poorly in histology, as relevant regions are often difficult to distinguish and localize using image-level supervision alone. To address these challenges, recent methods such as PixelCAM [44] have been specifically developed to improve weakly supervised localization in this setting. As such, histology provides a rigorous benchmark for evaluating the robustness and generalizability of WSOL approaches.

Quantitative results. Table 4 shows that TeD-Loc consistently outperforms all WSOL models, including the recent PixelCAM [44], on both localization and classification. On GlaS, classification is saturated for all methods (100.0%), so the meaningful comparison is localization, where TeD-Loc reaches 88.8% PxAP. On CAMELYON17, the gains are larger and consistent across all test centers. Compared to the recent method PixelCAM [44], TeD-Loc improves localization across all centers by +26.4%, +27.5%, +16.8%, +13.2%, and +31.4%, respectively. Averaged across the CAMELYON17 centers, this corresponds to an improvement of +23.1% in localization and +22.2% in classification over PixelCAM [44]. These gains indicate that the proposed method not only produces better activation maps, but also leads to more reliable classification under cross-center variability in histology.

To assess whether the gains of TeD-Loc are driven by the pretrained VLM, we compare zero-shot performance when using either the pathology-specialized CONCH initialization or an initialization pretrained on natural scene images. As we can observe in Table 5, using the pretrained model CONCH [43] did not impact the classification accuracy, with similar results on GlaS and CAMELYON17. For localization, CONCH provides slight improvements overall, with the largest gain observed on GlaS (+4.4% PxAP). Regarding CAMELYON17, the differences remain small and inconsistent across centers. These results highlight that the observed performance of TeD-Loc is mainly due to the proposed method rather than being explained by the encoder pretraining.

Qualitative results. Fig. 5 presents a visual comparison of TeD-Loc against standard WSOL baselines on the C17-4 *cancer* class. As shown in the figure, baseline CAM-based methods often activate irrelevant regions or only partial ROIs. TeD-Loc produces sharper and less noisy localization, reducing the false positives while preserving the relevant ROIs of the tissues.

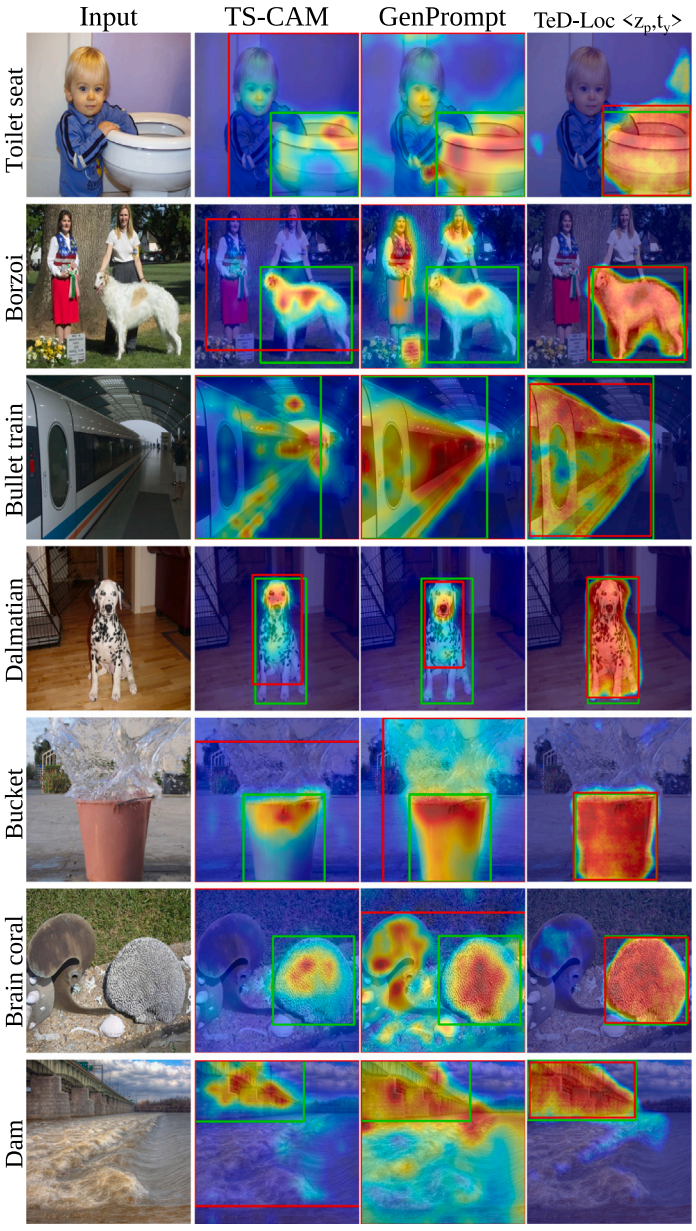


Fig. 4. Qualitative comparison on ILSVRC data. Localization maps are obtained via (patch, class) embeddings dot product: $\langle z_p, t_y \rangle$ where $z_p, \forall p \in \Omega$ are the patch embeddings. GenPrompt fails to localize objects in complex scenes, often because it relies on external classifiers to compute text embeddings. This can fail if the classifier makes mistakes. This dependence on class labels during inference highlights GenPrompt's vulnerability to localization errors. In contrast, TeD-Loc can localize objects in complex scenes. Here, green bboxes denote GT localization, while red bboxes represent predicted localizations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Localization (PxAP) and classification (CL) accuracy on the GlaS and CAMELYON17 center-wise test datasets.

Methods	GlaS		C17-0		C17-1		C17-2		C17-3		C17-4	
	PxAP (↑)	CL (↑)	PxAP (↑)	CL (↑)	PxAP (↑)	CL (↑)	PxAP (↑)	CL (↑)	PxAP (↑)	CL (↑)	PxAP (↑)	CL (↑)
DeepMIL [50] *(icml,2018)*	79.9	**100.0**	34.8	82.8	30.1	80.8	31.3	63.8	27.6	88.6	18.0	59.4
GradCAM++ [51] *(wacv,2018)*	76.8	**100.0**	21.9	72.1	22.2	66.3	29.8	80.8	32.4	77.6	21.4	59.7
LayerCAM [52] *(tip,2021)*	75.1	**100.0**	22.8	72.1	22.6	66.3	30.1	80.8	33.1	77.6	21.8	66.1
SAT [53] *(iccv,2023)*	65.9	**100.0**	20.6	64.5	17.5	72.7	27.7	65.6	20.4	52.6	18.9	49.6
PixelCAM [44] *(midl,2025)*	86.6	**100.0**	49.8	80.9	52.2	73.8	54.9	73.4	71.9	65.1	50.6	61.1
TeD-Loc	**88.8**	**100.0**	**76.2**	**93.5**	**79.7**	**95.3**	**71.7**	**89.1**	**85.1**	**94.4**	**82.0**	**92.9**

4.2.3. Ablations

Impact of losses. Table 6 shows that the combination of our proposed loss functions: knowledge-distillation loss ($\mathbf{L}_{\text{KD}}$), patch classifier loss ($\mathbf{P}_{\text{CL}}$) and global image classification loss ($\mathbf{I}_{\text{CL}}$), is essential for achieving state-of-the-art localization performance. Using only the main loss $\mathbf{L}_{\text{KD}}$, our method yields a MaxBoxAcc of 62.3%, MaxBoxAcc of 62.3%, as it focuses on learning FG embeddings without yielding explicit

Table 5
Zero-shot localization (PxAP) and classification (CL) accuracies of TeD-Loc with and without CONCH initialization on GlaS and CAMELYON17 center-wise test sets.

Methods	GlaS		C17-0		C17-1		C17-2		C17-3		C17-4	
	PxAP (↑)	CL (↑)	PxAP (↑)	CL (↑)	PxAP (↑)	CL (↑)	PxAP (↑)	CL (↑)	PxAP (↑)	CL (↑)	PxAP (↑)	CL (↑)
TeD-Loc wo/ CONCH [43]	49.3	53.7	16.1	50.0	14.7	50.0	27.2	50.0	20.9	50.0	19.2	50.0
TeD-Loc w/ CONCH [43]	53.5	53.7	16.7	50.0	14.8	50.0	28.0	50.0	19.8	50.0	19.4	50.0

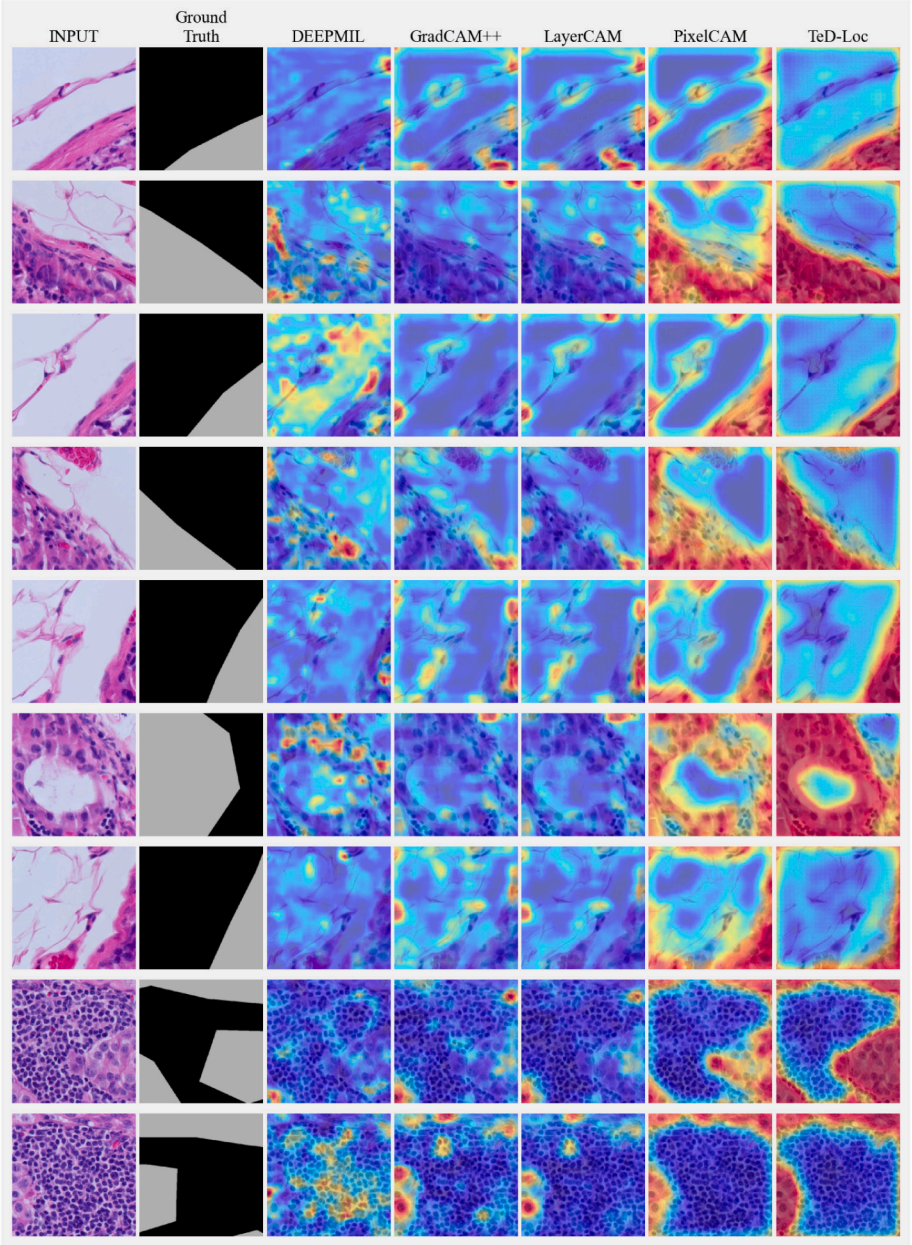


Fig. 5. Visualization of localization map obtained via TeD-Loc as compared to standard WSOL models on the C17-4 dataset for the class cancer.

localization, thereby explaining its limited performance. Adding patch classifier loss significantly improves accuracy to 95.4% by enhancing FG/BG separation. Finally, incorporating image class loss boosts the performance to 98.7%, emphasizing the importance of discriminative learning to distinguish between correct and incorrect class alignments. This demonstrates the effect of integrating these losses to achieve highly accurate weakly supervised object localization.

Hyperparameters sensitivity. Table 7 analyzes the sensitivity of TeD-Loc to the loss weights λ_1, λ_2, λ_3, which control the contribution of the knowledge-distillation, patch classification, and image classification losses, respectively. Results on GlaS show that TeD-Loc performs best under a balanced weighting of these objectives. In particular, the best performances are obtained if all lambdas are ≤ 1. Larger values generally lead to a decline in PxAP (see Table 7).

Impact of different CAM modules on TeD-Loc performance. Table 8 analyzes the impact of the CAM pseudo-labels source on C17-4. While the choice of CAM slightly affects performance, all different modules achieve strong results, with localization ranging from 78.3

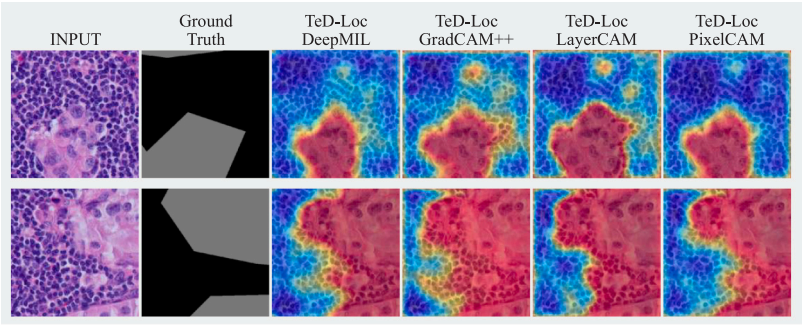


Fig. 6. Visualization of localization map obtained via TeD-Loc on the C17-4 dataset for the class cancer with different CAM modules (DeepMil [50], GradCAM++ [51], LayerCAM [52], PixelCAM [44]).

Table 6
Ablation study on the CUB dataset showing the impact of different loss combinations on MaxBoxAcc performance.

Losses	CUB (MaxBoxAcc)
$\mathbf{L}_{KD}$	62.3
$\mathbf{L}_{KD}$+$\mathbf{P}_{CL}$	95.4
$\mathbf{L}_{KD}$+$\mathbf{P}_{CL}$+$\mathbf{I}_{CL}$	**98.7**

Table 7
Ablation study of the loss weights on GlaS. Columns report the values tested for λ_1, λ_2, and λ_3.

Dataset	λ_1				λ_2				λ_3			
	0.5	1.0	1.5	2.0	0.5	1.0	1.5	2.0	0.5	1.0	1.5	2.0
GlaS	89.1	88.8	78.7	78.2	78.1	88.8	78.6	78.6	78.3	88.8	86.4	78.8

Table 8
Impact of different CAM-based pseudo-labels on TeD-Loc performance. Localization (PxAP) and classification (CL) accuracies on test set of C17-4 dataset.

Methods	C17-4	
	PxAP (↑)	CL (↑)
DeepMIL [50]	18.0	59.4
TeD-Loc w/ DeepMIL	**78.3**	**89.3**
GradCAM++ [51]	21.4	59.7
TeD-Loc w/ GradCAM++	**81.1**	**94.3**
LayerCAM [52]	21.8	66.1
TeD-Loc w/ LayerCAM	**80.4**	**93.9**
PixelCAM [44]	50.6	61.1
TeD-Loc w/ PixelCAM	**82.0**	**92.9**

to 82.0 PxAP and classification ranging from 89.3 to 94.3. These performances are higher than those of the corresponding standard WSOL models reported in Table 4, whose localization performance ranges from 18.0 to 50.6 PxAP on the center C17-4. Also, Fig. 6 shows that TeD-Loc produces qualitatively similar localization maps across different CAM modules. Although stronger CAM modules can yield slightly cleaner boundaries and fewer false-positive activations, the overall gain appears largely independent of the specific CAM module used.

4.3. Patch-level localization with text anchors vs. localization from $g(z_p)$

In the standard CLIP model [24], text class embedding is not necessarily correlated with the local vision patch embedding. To show this, we conduct the following experiment: consider the text embedding of the image class label y: t_y. Then, to localize this class object within the image, we perform a dot product across all patch embeddings: $\langle z_p, t_y \rangle$ where $z_p, \forall p \in \Omega$. High scores at location p should indicate the high likelihood of the object y presence at this location. The obtained score map is then considered as a CAM. We perform this experiment over three variants of CLIP model: Vanilla CLIP [24], SCLIP [26], and NACLIP [27], in addition to our method. The obtained results are presented in Table 9. Vanilla CLIP yielded poor results, confirming that class text embeddings are not necessarily correlated with the local patch embeddings, making them less useful for this task. This justifies using a gradient-based method over the dot product score between the global image embedding and a class text embedding in CLIP-ES [11]. However, we notice a greater improvement in localization for the more recent CLIP variants SCLIP [26], and NACLIP [27]. However, their Top-1, and Top-5 Loc are still low, indicating poor classification performance. On the other hand, our method achieves the highest performance over the three metrics, indicating better localization and classification scores over both datasets. This is the result of our text-to-patch distillation, which ensures that local patch embeddings are correlated with the class text embedding, allowing direct localization based on the text embedding. This equips our method with a secondary localization approach and the patch FG/BG classifier g. This second localization approach yielded relatively better performance than when using g. In addition to this, we evaluate the localization map extracted from module $g(z_p)$ of our method, which performs slightly lower than the results obtained via $\langle z_p, t_y \rangle$ as shown in Table 9. Furthermore, Fig. 7 visualizes this localization strategy of different CLIP variants and our method.

Impact of Orthogonalization on Performance. Table 10 shows the impact of orthogonalization of class text embedding in our method for both tasks: classification and localization (obtained via $g(z_p)$). These results suggest that original class text embeddings of CLIP [24] are not adequate to perform discriminative learning, as these embeddings overlap, as shown visually in the main paper (Fig. 3). However, their orthogonalization allows better separation of these embeddings, making them more suitable for classification tasks. This also positively affects localization as well in our method, since both tasks are strongly related by design.

4.4. Failures cases

TeD-Loc provides clear advantages and significantly improves localization performance over standard WSOL models, but it still fails in some challenging cases for both natural and medical images. For natural images, Fig. 8(a) illustrates several typical failure modes. In the "coral reef" example, the presence of fish within the reef scene appears to mislead the model into partially associating the target class with the fish, which results in activations over both the fish and the reef rather than a more precise localization of the coral region. In the "bicycle tire" example, the activation extends beyond the target part and covers a broader portion of the wheel, suggesting limited spatial precision for fine-grained categories. In the "desktop computer" example, the model

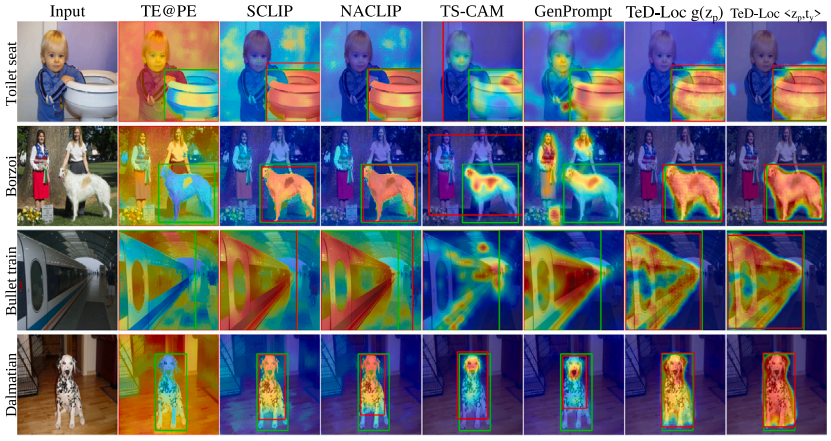


Fig. 7. Visualization of localization map obtained via $g(z_p)$ as compared to (patch, class) embeddings dot product: $\langle z_p, t_y \rangle$ where $z_p, \forall p \in \Omega$ is the patch embeddings, and y is the true image class over different variants of CLIP, and our method. TE@PE is the vanilla CLIP [24] where TE is the text embedding, and PE is the patch embedding.

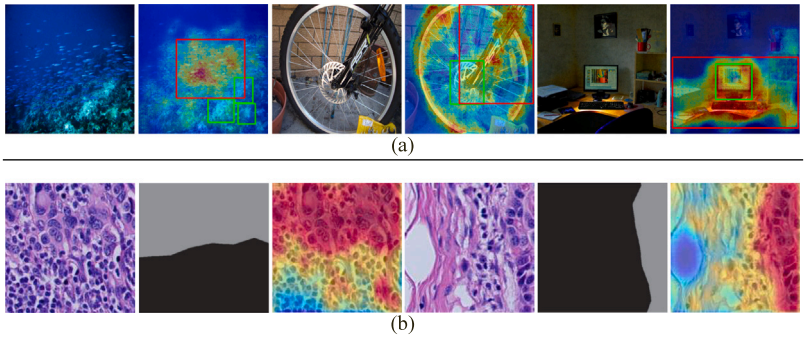


Fig. 8. Failure cases of TeD-Loc in localizing ROIs. Row (a) shows 3 examples from natural images (ILSVRC) where GT annotations are bboxes and localization is evaluated via MaxBoxAcc. Row (b) shows 3 examples from histology images (CAMELYON17), where GT annotations are pixel-level segmentation masks, and localization is evaluated via PxAP.

Table 9
Localization performance of maps obtained via $g(z_p)$ and (patch, class) embeddings dot product: $\langle z_p, t_y \rangle$ where $z_p, \forall p \in \Omega$ is the patch embeddings, and y is the true image class. We report localization performance (MaxBoxAcc, **Top-1 Loc**, **Top-5 Loc**) using different variants of CLIP and our method.

	CUB			ILSVRC		
	MaxBoxAcc	Top-1 Loc	Top-5 Loc	MaxBoxAcc	Top-1 Loc	Top-5 Loc
Vanilla CLIP [24]	18.8	8.9	15.2	41.1	26.6	37.0
SCLIP (eccv'24) [26]	85.8	14.4	37.1	70.4	33.9	55.1
NACLIP (wacv'25) [27]	80.8	12.0	32.3	71.7	28.6	49.4
TeD-Loc w/ $g(z_p)$	**75.6**	**70.0**	**75.1**	**98.7**	**91.7**	**97.6**
TeD-Loc w/ $\langle z_p, t_y \rangle$	**77.2**	**71.8**	**76.7**	**98.7**	**92.0**	**97.5**

Table 10
Impact of class text embeddings (text anchors) orthogonalization over localization and classification performance in our method over CUB dataset.

TeD-Loc	CUB			
w/o orthogonalization, with default anchors and $g(z_p)$	97.7	56.0	54.8	85.8
w/ orthogonalization, with $g(z_p)$	**98.7**	**93.0**	**91.7**	**97.6**

also responds to the surrounding desk area, indicating that contextual cues may dominate when the object occupies only a limited portion of the image. More generally, these examples show that, although TeD-Loc identifies semantically relevant regions, the predicted localization can remain spatially coarse and may over-extend beyond the true object boundaries.

A similar limitation is observed in histology images, as shown in Fig. 8(b). In these examples, TeD-Loc correctly assigns high confidence to diagnostically relevant tissue regions, but the activation map may still spread into adjacent background. This suggests that the method captures the most informative regions, yet does not always delineate the ROI sharply. Such behavior is consistent with the intrinsic difficulty of histology WSOL, where boundaries between relevant and non-relevant tissue are often gradual, local morphology can be highly heterogeneous, and the amount of annotated training data is limited [40]. These observations indicate that, while text-guided alignment improves semantic localization, additional spatial regularization or refinement strategies may be required to achieve more precise boundary delineation in challenging settings characterized by limited data and strong foreground–background similarity.

5. Conclusion

We have introduced TeD-Loc, a novel WSOL approach that integrates textual and visual modalities by transferring knowledge directly from CLIP's text embeddings to our patch embedding module. This alignment enables our model to localize objects at the patch-level while simultaneously performing global image classification. Our results show that language-guided visual alignment can effectively benefit WSOL tasks and further suggest that stronger localization can directly benefit image-level recognition. Additionally, the proposed QR-based text embedding orthogonalization improves discriminability for semantically similar classes. Experiments on CUB, ILSVRC, and two histopathology benchmarks (GlaS and CAMELYON17) show consistent improvements over baseline methods. For instance, our method improves up to 31% PxAP on histopathology datasets, while offering more efficient inference than GenPrompt. Regarding limitations, although TeD-Loc improves localization performance, the qualitative results show that it can still fail in challenging cases, either by missing parts of the relevant ROI or by producing over-activations. These failures suggest that semantic text-guided alignment alone is not always sufficient to guarantee precise spatial localization, and that additional spatial regularization or refinement strategies could further improve localization quality. While the current study focuses on single-object localization benchmarks, the class-conditioned design of TeD-Loc suggests that it could naturally extend to multi-class or multi-instance settings. However, this potential remains to be validated experimentally. Finally, the behavior of CLIP-based text guidance in highly specialized domains, such as remote sensing or industrial inspection, beyond those studied here remains to be investigated. We believe TeD-Loc opens new directions for WSOL by distilling class-level semantic knowledge from the CLIP text encoder into patch representations using only image-level supervision. The text distillation paradigm could be plugged into the backbones. Future work could explore extending TeD-Loc to open-vocabulary localization settings and to multi-label scenarios, and investigate domain adaptation to improve generalization across histopathology datasets.

CRediT authorship contribution statement

Shakeeb Murtaza: Writing – original draft, Validation, Methodology, Conceptualization. **Soufiane Belharbi:** Writing – review & editing, Supervision, Conceptualization. **Alexis Guichemerre:** Writing – review & editing, Validation, Formal analysis. **Marco Pedersoli:** Writing – review & editing, Supervision, Funding acquisition. **Eric Granger:** Writing – review & editing, Supervision, Investigation, Conceptualization.

AI Declaration

Generative AI tools were used solely for grammar correction, formatting, and spelling.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada. We also thank the Digital Research Alliance of Canada for the use of their computing resources.

Data availability

I am using public datasets, and our code is available on github.

References

[1] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: CVPR, 2016.
[2] S. Belharbi, A. Sarraf, M. Pedersoli, I.B. Ayed, L. McCaffrey, E. Granger, F-CAM: Full resolution class activation maps via guided parametric upscaling, in: WACV, 2022.
[3] P. Wu, W. Zhai, Y. Cao, Background activation suppression for weakly supervised object localization, in: CVPR, 2022.
[4] H. Xue, C. Liu, F. Wan, J. Jiao, X. Ji, Q. Ye, DANet: Divergent activation for weakly supervised object localization, in: CVPR, 2019.
[5] X. Zhang, Y. Wei, J. Feng, Y. Yang, T. Huang, Adversarial complementary learning for weakly supervised object localization, in: CVPR, 2018.
[6] J. Choe, H. Shim, Attention-based dropout layer for weakly supervised object localization, in: CVPR, 2019.
[7] J. Choe, D. Han, S. Yun, J.-W. Ha, S.J. Oh, H. Shim, Region-based dropout with attention prior for weakly supervised object localization, Pattern Recognit. 116 (2021).
[8] X. Zhang, Y. Wei, G. Kang, Y. Yang, T. Huang, Self-produced guidance for weakly-supervised object localization, in: ECCV, 2018.
[9] S. Murtaza, S. Belharbi, M. Pedersoli, A. Sarraf, E. Granger, Discriminative sampling of proposals in self-supervised transformers for weakly supervised object localization, in: WACV Workshop, 2023.
[10] W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, B. Zhou, Q. Ye, TS-CAM: Token semantic coupled attention map for weakly supervised object localization, in: ICCV, 2021.
[11] Y. Lin, M. Chen, W. Wang, B. Wu, K. Li, B. Lin, H. Liu, X. He, CLIP is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation, in: CVPR, 2023.
[12] J. Wang, G. Kang, Learn to rectify the bias of CLIP for unsupervised semantic segmentation, in: CVPR, 2024.
[13] Y. Zhao, Q. Ye, W. Wu, C. Shen, F. Wan, Generative prompt model for weakly supervised object localization, in: CVPR, 2023.
[14] W. Gander, Algorithms for the QR-decomposition, 1980.
[15] M.-A. Carbonneau, V. Cheplygina, E. Granger, G. Gagnon, Multiple instance learning: A survey of problem characteristics and applications, Pattern Recognit. (2018).
[16] W. Hui, C. Tan, G. Gu, Y. Zhao, Gradient-based refined class activation map for weakly supervised object localization, Pattern Recognit. 128 (2022).
[17] K. Singh, Y. Lee, Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization, in: ICCV, 2017.
[18] S. Yun, D. Han, S.J. Oh, S. Chun, J. Choe, Y. Yoo, CutMix: Regularization strategy to train strong classifiers with localizable features, in: ICCV, 2019, pp. 6023–6032.
[19] X. Pan, Y. Gao, Z. Lin, F. Tang, W. Dong, H. Yuan, F. Huang, C. Xu, Unveiling the potential of structure preserving for weakly supervised object localization, in: CVPR, 2021.
[20] C. Zhang, Y. Cao, J. Wu, Rethinking the route towards weakly supervised object localization, in: CVPR, 2020.
[21] Z. Chen, Y. Shen, L. Cao, S. Zhang, R. Ji, Clip-driven transformer for weakly supervised object localization, TPAMI (2025).
[22] S. Belharbi, S. Murtaza, M. Pedersoli, I. Ben Ayed, L. McCaffrey, E. Granger, CoLo-CAM: Class activation mapping for object co-localization in weakly-labeled unconstrained videos, Pattern Recognit. 162 (2025).
[23] J. Xie, Z. Luo, R. Li, Y. Huang, H. Liu, Y. Li, Y. Zheng, Y. Zhang, L. Shen, M.Z. Shou, Open-world weakly-supervised object localization, Pattern Recognit. 169 (2026).
[24] A. Radford, J. Kim, C. Hallacy, et al., Learning transferable visual models from natural language supervision, in: ICML, 2021.
[25] J. Xie, X. Hou, K. Ye, L. Shen, CLIMS: Cross language image matching for weakly supervised semantic segmentation, in: CVPR, 2022.
[26] F. Wang, J. Mei, A. Yuille, SCLIP: Rethinking self-attention for dense vision-language inference, in: ECCV, 2024.
[27] S. Hajimiri, I. Ayed, J. Dolz, Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation, in: WACV, 2025.
[28] L. Zhu, X. Wang, J. Feng, T. Cheng, Y. Li, B. Jiang, D. Zhang, J. Han, Weakclip: Adapting clip for weakly-supervised semantic segmentation, IJCV (2025).
[29] X. Yang, X. Gong, Foundation model assisted weakly supervised semantic segmentation, in: WACV, 2024.
[30] A. Kirillov, E. Mintun, N. Ravi, et al., Segment anything, in: ICCV, 2023.
[31] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, Y. Cao, EVA: Exploring the limits of masked visual representation learning at scale, in: CVPR, 2023.
[32] J. Choe, S. Oh, S. Lee, S. Chun, Z. Akata, H. Shim, Evaluating weakly supervised object localization methods right, in: CVPR, 2020.
[33] S. Murtaza, S. Belharbi, M. Pedersoli, E. Granger, A realistic protocol for evaluation of weakly supervised object localization, in: WACV, 2025.
[34] J. Rony, S. Belharbi, J. Dolz, I.B. Ayed, L. McCaffrey, E. Granger, Deep weakly-supervised learning methods for classification and localization in histology images: A survey, MLBI (2023).

[35] S. Murtaza, S. Belharbi, M. Pedersoli, A. Sarraf, E. Granger, DIPS: Discriminative pseudo-label sampling with self-supervised transformers for weakly supervised object localization, IVC J. (2023).
[36] N. Otsu, A threshold selection method from gray-level histograms, TSMC (1979).
[37] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: CVPR, 2009.
[38] C. Wah, S. Branson, W. Steve, P. Peter, S. Belongie, The Caltech-UCSD Birds-200–2011 Dataset, Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011.
[39] B. Recht, R. Roelofs, L. Schmidt, V. Shankar, Do imagenet classifiers generalize to imagenet? in: ICML, 2019.
[40] A. Guichemerre, B. Karimian, S. Belharbi, N. Gillet, N. Thome, P. Shamsolmoali, M. Shateri, L. McCaffrey, E. Granger, Adaptation of weakly supervised localization in histopathology by debiasing predictions, 2026, CoRR abs/2603.21309.
[41] T. Thomas, B. Alexe, V. Ferrari, Weakly supervised localization and learning with generic knowledge, IJCV (2012).
[42] A. Guichemerre, S. Belharbi, T. Mayet, et al., Source-free domain adaptation of weakly-supervised object localization models for histology, in: CVPR Workshop, 2024.
[43] M.Y. Lu, B. Chen, D.F. Williamson, R.J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L.P. Le, G. Gerber, et al., A visual-language foundation model for computational pathology, Nature Med. (2024).
[44] A. Guichemerre, S. Belharbi, M. Shateri, et al., PixelCAM: Pixel class activation mapping for histology image classification and ROI localization, in: MIDL, 2025.
[45] H. Bai, R. Zhang, J. Wang, X. Wan, Weakly supervised object localization via transformer with implicit spatial calibration, ECCV (2022).
[46] Z. Chen, C. Wang, Y. Wang, G. Jiang, Y. Shen, Y. Tai, C. Wang, W. Zhang, L. Cao, LCTR: on awakening the local continuity of transformer for weakly supervised object localization, in: AAAI, 2022, pp. 410–418.
[47] J. Xie, J. Xiang, J. Chen, X. Hou, X. Zhao, L. Shen, C2AM: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation, in: CVPR, 2022.
[48] Z. Chen, J. Ding, L. Cao, Y. Shen, S. Zhang, G. Jiang, R. Ji, Category-aware allocation transformer for weakly supervised object localization, in: ICCV, 2023.
[49] L. Zhu, Q. She, Q. Chen, Q. Ren, Y. Lu, Boosting weakly supervised object localization and segmentation with domain adaption, TPAMI (2024).
[50] M. Ilse, J.M. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: ICML, 2018.
[51] A. Chattopadhyay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks, in: WACV, 2018, pp. 839–847.
[52] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, Y. Wei, LayerCAM: Exploring hierarchical class activation maps for localization, TMI 30 (2021).
[53] P. Wu, W. Zhai, Y. Cao, J. Luo, Z.-J. Zha, Spatial-aware token for weakly supervised object localization, in: ICCV, 2023.
[54] Y. Cao, Y. Su, W. Wang, Y. Liu, Q. Wu, Semantic-constraint matching for transformer-based weakly supervised object localization, Pattern Recognit. 158 (2025).