CONFIDENTIAL EXTENDED ABSTRACT.
DO NOT DISTRIBUTE ANYWHERE.

ゼロショット異常検知のための学習可能ベクトルと階層構造への適応によるファインチューニング

七海 友康[1,a)]　都築 幸乃[1]　吉田 龍人[1]　大久保 順一[1]　藤井 純一郎[1]　山下 隆義[2]

概要

本研究では，ゼロショット異常検知のための CLIP のファインチューニングにおいて，学習対象と中間層特徴量に着目し，学習可能ベクトルを用いてモデルを軽量化しつつ性能を改善する手法を提案する．提案手法は，学習可能ベクトルでテキスト特徴量を代替することでテキストエンコーダを省略でき，モデルのパラメータを削減し，計算効率を向上させる．また，階層ごとに学習可能ベクトルを適用することで特徴表現の違いに適応し，さらに重みづけによって効果的な層を強調する．実験では，提案手法が少ないパラメータ数と高速な推論速度で，ピクセルレベルの評価において既存手法を上回る性能を達成することを示す．

1. はじめに

画像による異常検知は画像の正常・異常を画像レベルやピクセルレベルで判断するタスクであり，工業製品の欠陥検出をはじめとする様々な場面に適用される．近年，視覚と言語の特徴を統合した VLM (Vision and Language Model) が登場し，特に CLIP (Contrastive Language-Image Pre-training [12]) の汎化能力がさまざまな下流タスクで示されていることから，未知カテゴリの異常を検知するゼロショット異常検知が注目されている．WinCLIP [6] はその先駆けであり，CLIP のゼロショット画像分類フレームワークに基づいて異常を検知するアプローチを提案している．

CLIP を異常検知に適応させるためにモデル全体を再学習させる方法が考えられるが，これは非常に高い学習コストがかかる．小規模な学習で CLIP をゼロショット異常検知に適応させるためのアプローチとして，ターゲット以外のドメインのデータセットを用いた追加のパラメータの教師ありファインチューニングがある [2], [4], [15]．APRIL-GAN [4] は画像エンコーダの中間層におけるパッチレベルの特徴量を視覚言語空間にマッピングする追加の線形層を学習する．AnomalyCLIP [15] はプロンプト学習 [10], [13], [14] を取り入れ，オブジェクト非依存のプロンプト学習を導入する．また，CLIP のパラメータを固定させつつ，より画像の局所を重視した特徴量を獲得するための DPAM (Diagonally Prominent Attention Map) という機構を導入する．AdaCLIP [2] は学習可能トークンを CLIP の各エンコーダに組み込む手法 [7], [8] の有効性を利用し，入力画像から生成される動的トークンと静的トーク

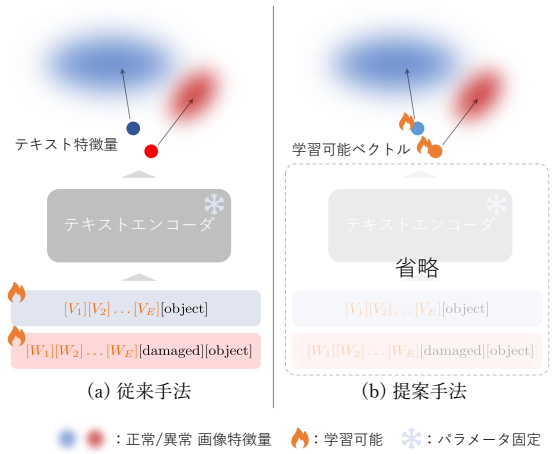

図 1: 学習対象に関する既存手法と提案手法の比較．(a) 既存手法はプロンプト学習で最適なテキスト特徴量の獲得を目指す一方，(b) 提案手法は学習可能ベクトルを直接最適化することでテキストエンコーダを省略可能である．

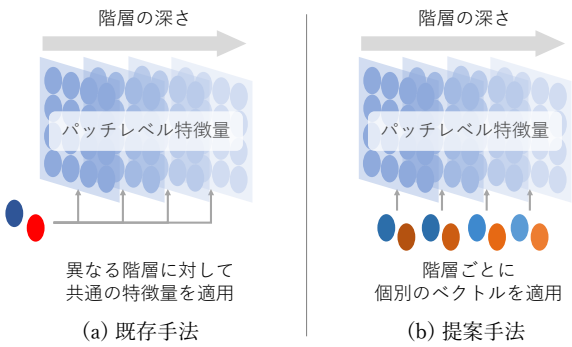

図 2: 異なる中間層へのアプローチに関する既存手法と提案手法の比較．(a) 既存手法は共通の特徴量を適用する一方，(b) 提案手法は各階層に個別のベクトルを適用する．

1　八千代エンジニヤリング株式会社
2　中部大学
a)　tm-nanaumi@yachiyo-eng.co.jp

ンを合わせた学習可能なトークンを各エンコーダに適用する．これらの既存研究について，次の 2 点に着目する．

(1) 正常・異常特徴量の獲得方法． AnomalyCLIP [15] は，オブジェクトに依存しない一般的な正常性と異常性を学習することを目的として，クラス名を含まない学習可能なテキストプロンプトの最適化によってテキスト特徴量を調整する (図 1 (a))．しかし，一般的な正常性と異常性を捉えるためには正常と異常に対応する 2 つのベクトルが最適化されれば十分であり，テキストエンコーダを用いることは必須ではないと考える．

(2) 中間層の特徴量に対するアプローチ． 既存手法 [2], [4], [15] では画像エンコーダの異なる中間層の特徴量に対して共通のテキスト特徴量を適用している (図 2 (a))．しかし，異なる中間層の特徴量が特徴空間を共有しているとは限らないため，共通のテキスト特徴量を適用することでモデルの性能が制限されている可能性がある．

上記 2 点について，本研究ではまず CLIP ベースのゼロショット異常検知手法を再考し，改善点を明確にする．その上で，画像エンコーダと追加の学習可能なパラメータのみで構成される，階層ごとの特徴量の違いに適応的なゼロショット異常検知手法を提案する．提案手法は，学習可能ベクトルのペアを導入し，直接最適化することでテキスト特徴量を代替する (図 1 (b))．これによりテキストエンコーダを省略できるため，モデルのパラメータを削減し，計算効率を向上させる．また，各階層に学習可能ベクトルを適用することで階層ごとの特徴表現の違いに適応し (図 2 (b))，さらに学習可能な重みパラメータによって異常検知に効果的な層の特徴を強調する．このアプローチにより異常度マップの精度を改善する．比較実験では，異常検知性能の定量的評価と異常度マップの可視化による定性的評価を行う．また，モデル内の学習可能なパラメータ数と総パラメータ数，推論速度でモデルの計算効率を評価する．

2. CLIP ベースのゼロショット異常検知

既存の CLIP ベースのゼロショット異常検知手法では，CLIP のゼロショット画像分類フレームワークに基づいて異常度スコアの計算を行う．CLIP の画像エンコーダは，画像 $\boldsymbol{I} \in \mathbb{R}^{H_{\mathrm{img}} \times W_{\mathrm{img}} \times 3}$ から特徴量を抽出する．ここで，最終出力のクラストークン $\boldsymbol{x} \in \mathbb{R}^D$ を画像レベル特徴量，l 層目のパッチトークン $\boldsymbol{x}^{(l)} \in \mathbb{R}^{H \times W \times D}$ をパッチレベル特徴量として扱う．テキストエンコーダは，ターゲットクラスセット $\mathcal{C}$ において，クラスを示す単語 $c \in \mathcal{C}$ と “A photo of a $[c]$.” のようなテンプレートを組み合わせたテキストから特徴量 $\boldsymbol{t}_c \in \mathbb{R}^D$ を抽出する．CLIP はこれらの特徴量から画像 $\boldsymbol{I}$ がクラス c に属する確率を次のように予測する:

$$P(\boldsymbol{x}, \boldsymbol{t}_c) = \frac{\exp(\mathrm{sim}(\boldsymbol{x}, \boldsymbol{t}_c)/\tau)}{\sum_{c' \in \mathcal{C}} \exp(\mathrm{sim}(\boldsymbol{x}, \boldsymbol{t}_{c'})/\tau)}, \tag{1}$$

ここで，$\mathrm{sim}(\cdot, \cdot)$ はコサイン類似度の計算であり，τ は温度スケーリングのハイパーパラメータである．このフレームワークを異常検知に適用する．$\mathcal{C} = \{\mathrm{n}, \mathrm{a}\}$ として，正常・異常を示すテキストから特徴量 $\boldsymbol{t}_{\mathrm{n}}, \boldsymbol{t}_{\mathrm{a}}$ をそれぞれ抽出し，異常度スコアを異常クラスに属する確率 $P(\boldsymbol{x}, \boldsymbol{t}_{\mathrm{a}})$ で算出する．パッチレベル特徴量 $\boldsymbol{x}^{(l)}$ に対しても同様の計算を適用することで，異常領域のセグメンテーションを実現する．

既存研究では，CLIP を異常検知に適応させるためにテストデータ以外のドメインのデータセットで教師ありファインチューニングを行う [2], [4], [15]．損失関数には Focal loss [9] と Dice loss [11] を用いて，計算された異常度スコアを正解ラベルの値に近づけるようにモデルを学習させる．式 (1) を踏まえると，異常度スコアは特徴量間の類似度に依存する．したがって，CLIP を用いたゼロショット異常検知の性能を高めるには画像特徴量かテキスト特徴量，あるいはその両方を調整することが必要であると考えられる．

3. 手法

3.1 既存手法の再考

画像認識タスクにおいては, プロンプト学習 [10], [13], [14] によって CLIP の汎化能力を活用する手法が開発されている．プロンプト学習は入力テキストプロンプトを学習可能な形式にし，各クラスに対してより適切なテキスト特徴量を獲得することで性能を向上させることを目指す．AnomalyCLIP はこのプロンプト学習手法の概念をゼロショット異常検知の文脈に適用しており，プロンプトをオブジェクトに依存しない形式で学習する．具体的には，学習可能なベクトル $[V_i], [W_i]$ $(i \in 1, \ldots, E)$ で入力テキストのテンプレートを構成し，クラス名を”object”で置き換えたプロンプト:

$$p_{\mathrm{n}} = [V_1][V_2] \ldots [V_E][\mathrm{object}] \tag{2}$$

$$p_{\mathrm{a}} = [W_1][W_2] \ldots [W_E][\mathrm{damaged}][\mathrm{object}] \tag{3}$$

を最適化することにより，一般的な正常性と異常性を捉えるテキスト特徴量を獲得することを目指す．しかし，本研究では図 1 に示すように，テキストプロンプトの改善ではなく，正常と異常を代表する特徴量の獲得方法に着目する．一般的な正常性と異常性を捉えるためには，各状態を代表するベクトルが獲得できれば十分であり，テキストエンコーダは必須ではないと考える．したがって，テキストエンコーダを使用しない手法により，性能を大幅に低下させることなく，モデルのパラメータ数の削減と推論速度の高速化を期待できる．

また，階層ごとの特徴表現の違いに対する適応性についても改善の余地がある．既存手法 [2], [4], [15] は画像レベル特徴量と各中間層のパッチレベル特徴量のいずれに対しても，単一の正常／異常テキスト特徴量のペアを共通して適用している．一方で，CLIP では各エンコーダの最終出力のみがアライメントするように学習されており，中間層

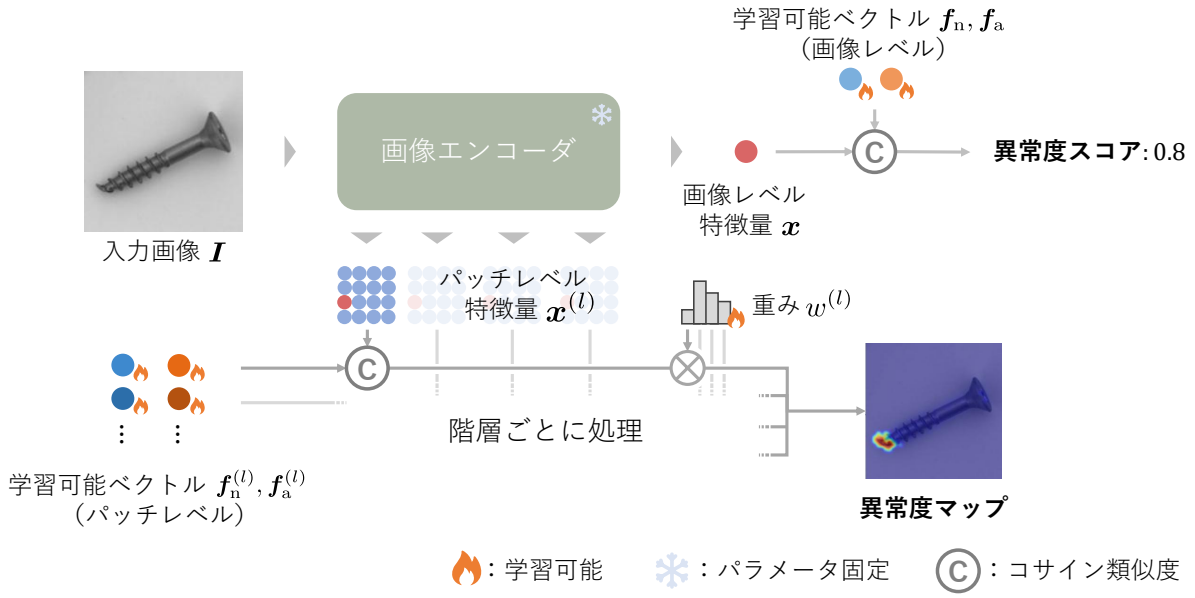

図 3: 提案手法の全体像．画像エンコーダから抽出された特徴量 $\boldsymbol{x}, \boldsymbol{x}^{(l)}$ に対して，学習可能ベクトルとの類似度によって異常度スコアを計算する．また，異なる中間層に対応する学習可能ベクトルのペアを導入し，それぞれを独立に最適化する．階層ごとに計算された異常度マップの重みつき総和を最終的な異常度マップとする．

では層の深さに応じて特徴表現の抽象度や空間的な構造が変化することが報告されている [3], [5]．そのため全層に共通のベクトルを適用するのは特徴表現のミスマッチが生じ，性能を制限する可能性がある．

3.2 提案手法

画像エンコーダのみを用いて，階層構造に対して適応的に追加のパラメータを最適化するゼロショット異常検知手法を提案する．手法の全体像を図 3 に示す．

画像エンコーダのみを用いたゼロショット異常検知． テキスト特徴量と同じサイズの学習可能なパラメータのペアを導入し，正常と異常をそれぞれ代表するベクトル $\boldsymbol{f}_{\mathrm{n}}, \boldsymbol{f}_{\mathrm{a}}$ とする．これらのベクトルはテキスト特徴量と同様に振る舞い，式 (1) に従って異常度スコアを $P(\boldsymbol{x}, \boldsymbol{f}_{\mathrm{a}})$ として算出する．このアプローチにより，テキストエンコーダを省略しながらも，プロンプト学習と同等の学習を実現できる．その結果，モデルのパラメータ数を削減し，推論速度を高速化することが期待できる．

階層の違いに適応したモデル設計． 異なる階層における特徴表現の違いに適応するために，画像レベル特徴量 $\boldsymbol{x}$ と各中間層のパッチレベル特徴量 $\boldsymbol{x}^{(l)}$ に対応するように学習可能ベクトルの組 $(\boldsymbol{f}_{\mathrm{n}}, \boldsymbol{f}_{\mathrm{a}}), (\boldsymbol{f}_{\mathrm{n}}^{(l)}, \boldsymbol{f}_{\mathrm{a}}^{(l)})$ を階層ごとに導入する．異なる階層に応じて個別の学習可能ベクトルを適用することで，独立にベクトルを最適化することができ，特徴表現の違いに対して柔軟に学習可能ベクトルを最適化する．さらに，異常検知のために効果的な層の異常度マップを強調するために，各層で計算された異常度マップに対して学習可能なパラメータ $w^{(l)}$ によって重要度重みづけを適用する．$w^{(l)}$ の最適化により，重要な層を強調する重みを獲得する．このアプローチにより，ピクセルレベルの異常検知性能を向上させることが期待できる．

4. 既存手法との比較実験

提案手法の有効性を検証するため，推論速度とピクセルレベルの性能について既存手法との比較実験を行う．

4.1 実験設定

バックボーンには事前学習済みの CLIP (ViT-L/14@336px[*1]) を用いる．ベースラインとして，既存の最先端手法である APRIL-GAN，AnomalyCLIP，AdaCLIP と性能を比較する．パッチレベル特徴量は，ベースラインに統一して $l \in \{6, 12, 18, 24\}$ 層目の特徴を利用する．データセットは MVTec AD [1] および VisA [16] を使用する．MVTec AD での評価には VisA を学習したモデルを，VisA での評価には MVTec AD を学習したモデルを使用する．評価は，異常検知結果の可視化による定性評価，異常検知の性能評価に用いられる評価指標，学習可能なパラメータ数，総パラメータ数，推論速度による定量評価を行う．性能評価はピクセルレベルの Area Under the Receiver Operating Characteristic (AUROC)，Average Precision (AP)，最適な閾値における F1 スコア (F1-max)，Per-Region Overlap (PRO) を用いて，カテゴリごとの性能の平均値を示す．画像レベルの AUROC, AP, F1-Max は参考値として示す．

[*1] https://huggingface.co/openai/clip-vit-large-patch14-336

表 1: MVTec AD および VisA データセットにおける性能の比較．単位は [%] で，大きい値ほど良い性能を示す．太字が最高性能，下線が 2 位の性能を示す．各特徴量の列は特徴量を調整するための手法を示す．

データセット	手法	画像特徴量	テキスト特徴量	ピクセルレベル				画像レベル		
				AUROC	AP	F1-Max	PRO	AUROC	AP	F1-Max
MVTec AD	APRIL-GAN	線形層	—	87.6	40.8	43.3	44.0	86.1	93.6	90.5
	AnomalyCLIP	DPAM	プロンプト学習	89.0	34.6	37.8	**81.8**	**90.3**	**95.7**	92.2
	AdaCLIP	線形層＋プロンプト学習	プロンプト学習	89.9	41.6	44.0	44.1	90.0	**95.7**	**92.4**
	提案手法	—	学習可能な特徴量	**91.3**	**42.9**	**44.4**	80.0	89.7	95.1	91.7
VisA	APRIL-GAN	線形層	—	94.2	25.7	32.3	86.8	78.0	81.4	78.7
	AnomalyCLIP	DPAM	プロンプト学習	94.3	20.9	26.9	86.8	80.6	83.7	79.5
	AdaCLIP	線形層＋プロンプト学習	プロンプト学習	95.8	**37.1**	31.5	51.3	**86.3**	83.5	**88.7**
	提案手法	—	学習可能な特徴量	**95.9**	29.8	**35.1**	**90.2**	83.3	**85.5**	81.7

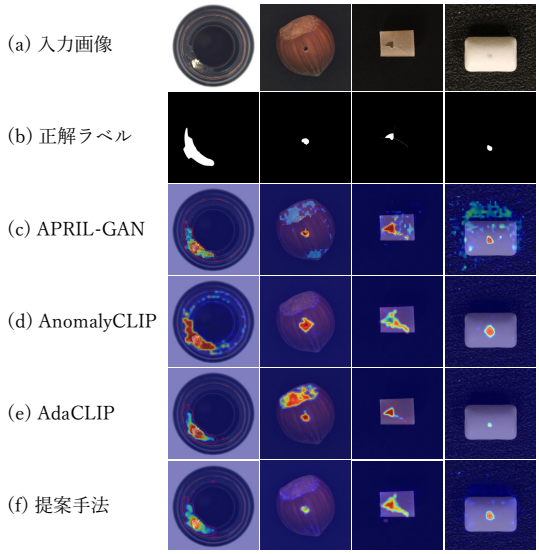

図 4: MVTec AD および VisA データセットにおける異常検知結果の可視化の比較．赤線は F1 スコアが最大となる時の閾値を示す．(f) の提案手法は (c)〜(e) の既存手法と同等以上の精度での異常検知が可能である．

4.2 実験結果

表 1 に性能の比較結果を示す．ピクセルレベルの性能について，MVTec AD では AUROC，AP，F1-Max で最高性能を達成し，PRO でも 2 位の性能を示す．VisA では AUROC，F1-Max，PRO で最も良い性能を示し，AP においても 2 位の性能を達成している．テキストエンコーダの省略は計算効率の向上にのみ寄与するため，階層ごとの特徴表現の違いに適応した設計がピクセルレベルの性能向上の要因であることが示唆される．画像レベルの性能については，提案手法はテキストエンコーダを使用しないにもかかわらず，既存手法と比較しても大幅には劣らない性能を維持している．図 4 に示した異常検知結果の可視化の比較においても，提案手法は既存手法と同等以上の検知結果を示すことが定性的に確認できる．

表 2 にモデルの学習可能パラメータ数と総パラメータ数，1 秒当たりの推論枚数を示す．既存手法が 3.2M〜10.7M の学習可能パラメータ数を必要とする一方で，提案手法はそれらと比較して学習可能パラメータを 99%以上削減した 7.7K パラメータしか必要としない．さらに，提案手法はテキストエンコーダを省略しているため，モデル全体のパラメータ数を少なくとも 29%削減することが可能である．また，既存手法は毎秒 1.8〜4.4 枚の画像を処理するが，提案手法はその 2 倍以上の毎秒 8.9 枚で最も良いスループットを達成している．表 1 の結果を踏まえると，提案手法はより少ないパラメータ数かつより早い推論速度で既存手法に匹敵する高い性能を達成することを示している．

表 2: モデルの学習可能なパラメータ数と総パラメータ数および推論速度の比較．最も良い数値を太字としている．

手法	学習可能パラメータ数	総パラメータ数	推論速度 [枚/秒]
APRIL-GAN	3.2M	431.9M	4.4
AnomalyCLIP	5.6M	433.5M	4.4
AdaCLIP	10.7M	439.4M	1.8
提案手法	**7.7K**	**304.3M**	**8.9**

5. おわりに

本研究では，CLIP の画像エンコーダと学習可能ベクトルで構成されるゼロショット異常検知手法を提案した．提案手法では，既存手法におけるテキストエンコーダを用いたプロンプト学習を省略し，学習可能ベクトルで代替することによって画像エンコーダのみでゼロショット異常検知を実現する．また，階層ごとに個別の学習可能ベクトルを適用し，計算された異常度マップに重みづけをすることで，特徴表現の違いに適応して学習可能ベクトルを最適化する．比較実験では，提案手法が既存手法と比較してモデルの学習可能パラメータ数を 99%以上，総パラメータ数を 29%以上削減し，かつ 2 倍以上の推論速度を実現しながら，ピクセルレベルの異常検知において既存手法に上回る性能を達成することを示した．画像レベルの性能改善手法については今後の課題である．

参考文献

[1] Bergmann, P., Fauser, M., Sattlegger, D. and Steger, C.: MVTec AD – A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection, *CVPR* (2019).

[2] Cao, Y., Zhang, J., Frittoli, L., Cheng, Y., Shen, W. and Boracchi, G.: AdaCLIP: Adapting CLIP with Hybrid Learnable Prompts for Zero-Shot Anomaly Detection, *ECCV* (2024).

[3] Chen, H., Lin, J., Chen, X., Fan, Y., Jin, X., Su, H., Dong, J., Fu, J. and Shen, X.: Rethinking Visual Layer Selection in Multimodal LLMs, *arXiv* (2025).

[4] Chen, X., Han, Y. and Zhang, J.: A Zero-/Few-Shot Anomaly Classification and Segmentation Method for CVPR 2023 VAND Workshop Challenge Tracks 1&2: 1st Place on Zero-shot AD and 4th Place on Few-shot AD, *CVPRW* (2023).

[5] Gandelsman, Y., Efros, A. A. and Steinhardt, J.: Interpreting the Second-Order Effects of Neurons in CLIP, *ICLR* (2025).

[6] Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A. and Dabeer, O.: WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation, *CVPR* (2023).

[7] Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B. and Lim, S.-N.: Visual Prompt Tuning, *ECCV* (2022).

[8] Khattak, M. U., Rasheed, H., Maaz, M., Khan, S. and Khan, F. S.: MaPLe: Multi-modal Prompt Learning, *CVPR* (2023).

[9] Lin, T., Goyal, P., Girshick, R. B., He, K. and Dollár, P.: Focal Loss for Dense Object Detection, *ICCV* (2017).

[10] Manli, S., Weili, N., De-An, H., Zhiding, Y., Tom, G., Anima, A. and Chaowei, X.: Test-Time Prompt Tuning for Zero-shot Generalization in Vision-Language Models, *NeurIPS* (2022).

[11] Milletari, F., Navab, N. and Ahmadi, S.-A.: V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation, *3DV* (2016).

[12] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision, *ICML* (2021).

[13] Zhou, K., Yang, J., Loy, C. C. and Liu, Z.: Conditional Prompt Learning for Vision-Language Models, *CVPR* (2022).

[14] Zhou, K., Yang, J., Loy, C. C. and Liu, Z.: Learning to Prompt for Vision-Language Models, *IJCV* (2022).

[15] Zhou, Q., Pang, G., Tian, Y., He, S. and Chen, J.: AnomalyCLIP: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection, *ICLR* (2024).

[16] Zou, Y., Jeong, J., Pemula, L., Zhang, D. and Dabeer, O.: SPot-the-Difference Self-Supervised Pre-training for Anomaly Detection and Segmentation, *ECCV* (2022).