پرسش ۱: DetailCLIP (۱۰۰ نمره)

۱.۱ یادگیری معلم-شاگردی در سطح بخش های تصویر (Patch-Level Self-Distillation)

(آ) این رویکرد سلسله‌مراتبی با ایجاد یک فرآیند یادگیری چندمقیاسی، به حفظ جزئیات تصد. در این روش، بخش های کوچک‌تر (شاگردان) از بخش های بزرگ‌تر (معلمان) یاد می‌گیرند. مدل معلم با پردازش بخش های بزرگ، زمینه کلی (Context) تصویر را فرا می‌گیرد، در حالی که مدل شاگرد بر روی بخش های کوچک‌تر تمرکز کرده تا جزئیات موضعی را استخراج کند. این انتقال دانش از نواحی بزرگ به نواحی کوچک، از نابودی ویژگی‌های ظریف که معمولاً هنگام پردازش تصویر تنها در یک مقیاس رخ می‌دهد، جلوگیری می‌کند.

(ب) مدل‌های دسته‌بندی سنتی محدودیت‌های زیر را دارند که این تکنیک به رفع آن‌ها کمک می‌کند:

- **از دست دادن اطلاعات در لایه‌های** Pooling**:** مدل‌های سنتی از عملیات ادغام استفاده می‌کنند که اطلاعات مکانی را دور می‌ریزند.
- **استخراج ویژگی‌های نادقیق:** Global Average Pooling اغلب جزئیات ریز را نادیده می‌گیرد.
- **مشکلات** Scale Invariance**:** مدل‌ها در برخورد با اشیاء در مقیاس‌های مختلف مشکل دارند.

یادگیری معلم-شاگردی در سطح تیکه با حفظ بازنمایی‌های چندمقیاسی و جزئیات مکانی از طریق انتقال دانش، بر این محدودیت‌ها غلبه می‌کند.

۲.۱ فیلترسازی مبتنی بر توجه (Attention-Based Filtering)

(آ) مدل با معیارهای زیر تصمیم می‌گیرد کدام نواحی تصویر را در اولویت قرار دهد:

- **وزن‌های** Attention**:** نمره توجه بالاتر نشان‌دهنده اهمیت بیشتر ناحیه است.
- **ارتباط** Semantic Relevance**:** نواحی حاوی اشیاء مورد علاقه اولویت بیشتری دریافت می‌کنند.
- **اهمیت** Contextual Importance**:** مناطقی که اطلاعات زمینه‌ای برای درک صحنه فراهم می‌کنند.
- **تمایز ویژگی (**Feature Distinctiveness**):** نواحی دارای ویژگی‌های منحصر به فرد یا متمایز، توجه بیشتری جلب می‌کنند.
- **ارتباط خاص وظیفه (**Task-Specific Relevance**):** برای وظایفی مانند Segmentation، نواحی مرزی و مراکز اشیاء اولویت‌بندی می‌شوند.

(ب) این عمل (انتخاب توجه) تأثیرات زیر را بر کیفیت تحلیل دارد:

- **کاهش پیچیدگی محاسباتی:** متمرکز شدن فقط روی نواحی مرتبط، محاسبات را ذخیره می‌کند.
- **بهبود دقت:** حذف نویز از مناطق غیرمهم، دقت را افزایش می‌دهد.
- **افزایش** Robustness**:** مدل را در برابر شلوغی پس‌زمینه کمتر آسیب‌پذیر می‌کند.
- **استخراج ویژگی بهتر:** متمرکز شدن بر نواحی از نظر معنایی مهم، یادگیری ویژگی را بهبود می‌بخشد.
- **همگرایی سریع‌تر:** مدل با تمرکز بر اطلاعات مرتبط، کارآمدتر یاد می‌گیرد.

۳.۱ بازسازی در سطح پیکسل (Pixel-Level Reconstruction)

(آ) توانایی افزایش وضوح ورودی‌های کم‌کیفیت به دلایل زیر در کاربردهای دنیای واقعی ارزشمند است:

- **محدودیت‌های سنسور:** دوربین‌ها و دستگاه‌ها اغلب تصاویر با وضوح پایین ثبت می‌کنند.
- **ملاحظات هزینه:** تجهیزات با وضوح بسیار بالا گران هستند.
- **کارایی انتقال:** وضوح پایین انتقال داده را سریع‌تر می‌کند.

(ب) انواع اشیاء و صحنه‌هایی که بیشتر از این قابلیت بهره می‌برند عبارتند از:

- **تصویربرداری پزشکی:** ساختار سلول‌ها، جزئیات بافت‌ها
- **تحلیل پزشکی قانونی (Forensic Analysis):** اثرانگشت، جزئیات اسناد
- **نمونه‌های بیولوژیکی:** ارگانیسم‌های میکروسکوپی، ساختارهای سلولی

۴.۱ ادغام مؤلفه‌ها

(آ) این سه مؤلفه به روش‌های زیر یکدیگر را تکمیل کرده و یک مدل قدرتمند می‌سازند:

- Self-Distillation: جزئیات ریز را حفظ می‌کند.
- Token Removal: پردازش را با متمرکز کردن منابع محاسباتی بهینه می‌کند.
- Pixel Reconstruction: کیفیت خروجی را از ورودی‌های با وضوح پایین افزایش می‌دهد.

(ب) در صورت حذف یکی از این تکنیک‌ها، ضعف‌های زیر ممکن است ایجاد شود:

- **بدون** Self-Distillation: از دست دادن جزئیات ریز، کاهش دقت Segmentation
- **بدون** Token Removal: افزایش هزینه محاسباتی، پردازش کندتر
- **بدون** Reconstruction: عملکرد ضعیف روی ورودی‌های کم‌کیفیت، کاربردپذیری محدود

پرسش ۲: یادگیری خودنظارتی (۵۰ نمره)

(الف) SimCLR به نمونه‌های نامشابه (Negative Samples) نیاز دارد تا از مشکل Collapsing Solution جلوگیری کند، جایی که همه بازنمایی‌ها به یک نقطه همگرا می‌شوند. نمونه‌های منفی سیگنال‌های Contrastive ارائه می‌دهند که با دور کردن نمونه‌های نامشابه از هم و نزدیک کردن نمونه‌های مشابه به هم، به مدل در یادگیری بازنمایی‌های معنادار کمک می‌کنند. بدون نمونه‌های منفی، مدل ممکن است به راه‌حل‌های پیش‌پاافتاده‌ای برسد که در آن همه بازنمایی‌ها یکسان هستند و هدف یادگیری ویژگی‌های متمایزکننده را نقض می‌کند.

(ب) دلیل اینکه این مشکل در روش‌هایی مانند BYOL رخ نمی‌دهد، معماری خاص آن است که از مکانیسم‌های زیر استفاده می‌کند:

- **معماری نامتقارن (Asymmetric Architecture):** پارامترهای شبکه متفاوت برای شبکه‌های Online و Target
- Momentum Encoder: شبکه Target به آرامی و با استفاده از Exponential Moving Average به روز می‌شود.

- Stop-Gradient: از دریافت گرادیان توسط شبکه Target جلوگیری می‌کند.
- Predictor Network**:** یک MLP اضافی که بازنمایی‌های Target را پیش‌بینی می‌کند.

این معماری یک پویایی یادگیری پایدار ایجاد می‌کند که در آن شبکه Online مجبور است بازنمایی‌های شبکه Target را پیش‌بینی کند و از راه‌حل‌های پیش‌پاافتاده جلوگیری می‌نماید.

(ج) دلیل مقاومت بیشتر BYOL در انتخاب هایپرپارامترها به عوامل زیر مربوط می‌شود:

- **عدم نیاز به** Negative Sampling**:** حساسیت به اندازه Batch Size را از بین می‌برد (روش‌های Contrastive به Batch Size بزرگ برای نمونه‌های منفی کافی نیاز دارند).
- **بهینه‌سازی پایدار:** Momentum Encoder اهداف پایداری برای یادگیری فراهم می‌کند.
- **کاهش خطر** Mode Collapse**:** معماری Asymmetric به طور طبیعی از فروپاشی جلوگیری می‌کند.
- **بهبود** Transformation Invariance**:** بدون تکیه بر تقابل منفی، بازنمایی‌های قوی‌تری یاد می‌گیرد.

(د) دلیل این استراتژی (نزدیک کردن Global Crops و دور کردن Local Crops) این است که:

- **یادگیری ویژگی‌های سلسله‌مراتبی:** Global Crops ساختار کلی را ثبت می‌کنند در حالی که Local Crops روی جزئیات تمرکز می‌کنند.
- **جلوگیری از** Over-Smoothing**:** از همگرایی همه بازنمایی‌ها به یک نقطه واحد جلوگیری می‌کند.
- **درک چندمقیاسی (**Multi-Scale Understanding**):** مدل را تشویق می‌کند تا هم زمینه کلی و هم جزئیات موضعی را درک کند.
- **یادگیری بازنمایی بهتر:** ویژگی‌های اطلاعاتی‌تر و متمایزکننده‌تری ایجاد می‌کند.

این رویکرد به‌ویژه در روش DINO (Distillation Without Labels) برای یادگیری خودنظارتی مؤثر است.