

Supplementary Materials: Hunting Blemishes: Language-guided High-fidelity Face Retouching Transformer with Limited Paired Data

Anonymous Authors

1 ADDITIONAL RESULTS

1.1 Blemish Detection

We propose a Language-guided Blemish Removal Transformer for automatically retouching face images, which is referred to as LangBRT. One of our design elements is the Text-prompted Blemish Detection module (TBD), which utilizes the prior knowledge encapsulated in a pre-trained vision-language model. Based on the text descriptions of facial blemishes, TBD is able to detect prompt-specific blemishes, which are used to guide the retouching process. In Figure 1, we visualize the blemish detection results for the cases where the textual prompts are “forehead pimples / reflection / messy hair”, “dark circle under the eyes”, “blemish on the skin” and so on. One can observe that the combinations of blemish prompts are useful for detecting more blemishes. We further perform a visual comparison between LangBRT and the state-of-the-art face retouching method, BPFRe. As shown in Figure 2, LangBRT outperforms BPFRe in blemish detection, since the heat maps of LangBRT indicate the blemish regions, which are more consistent with the manual retouching regions.

1.2 Comparison to State-of-the-arts

To highlight the superiority of our proposed LangBRT with both limited(1%) and full(1%) paired training data, we perform visual comparison to the main competing methods: ABPN and BPFRe on in-the-wild data. A number of representative retouching images of the methods are shown in Figure 3 and Figure 4. One can find that LangBRT achieves better retouching results than the main competing methods both with 1% and 100% FFHQ training data,

while at the same time preserving the non-blemish content of the input face images.

Further, to demonstrate the effectiveness of LangBRT, we present high-resolution examples for reflection and dark circle(Figure 5), messy hair and pimples(Figure 6). This confirms the significant advantages of our LangBRT in handling various types of blemishes.

2 SUMMARY OF LANGBRT

LangBRT consists of the following learnable components: the image encoder E , latent transformer T , decoder D and discriminator S , and their network architectures are presented in Tables 1, 2, 3 and 4, respectively. The constituent networks are jointly optimized from scratch, and the training procedure is summarized in Algorithm 1.

3 ETHICAL STATEMENT

This study on facial blemish removal adheres to strict ethical guidelines. Participants were fully informed about the study’s purpose and provided voluntary consent. We prioritize participant privacy, protecting their data through secure storage and access restrictions. The proposed retouching technique aims to enhance natural attributes while ensuring transparency and avoiding deceptive alterations. We acknowledge potential biases and limitations, emphasizing ongoing research to promote fair and inclusive application. Overall, we are committed to conducting this study ethically and contributing to responsible face retouching practices. The use of the proposed model for harmful purposes is strongly discouraged.

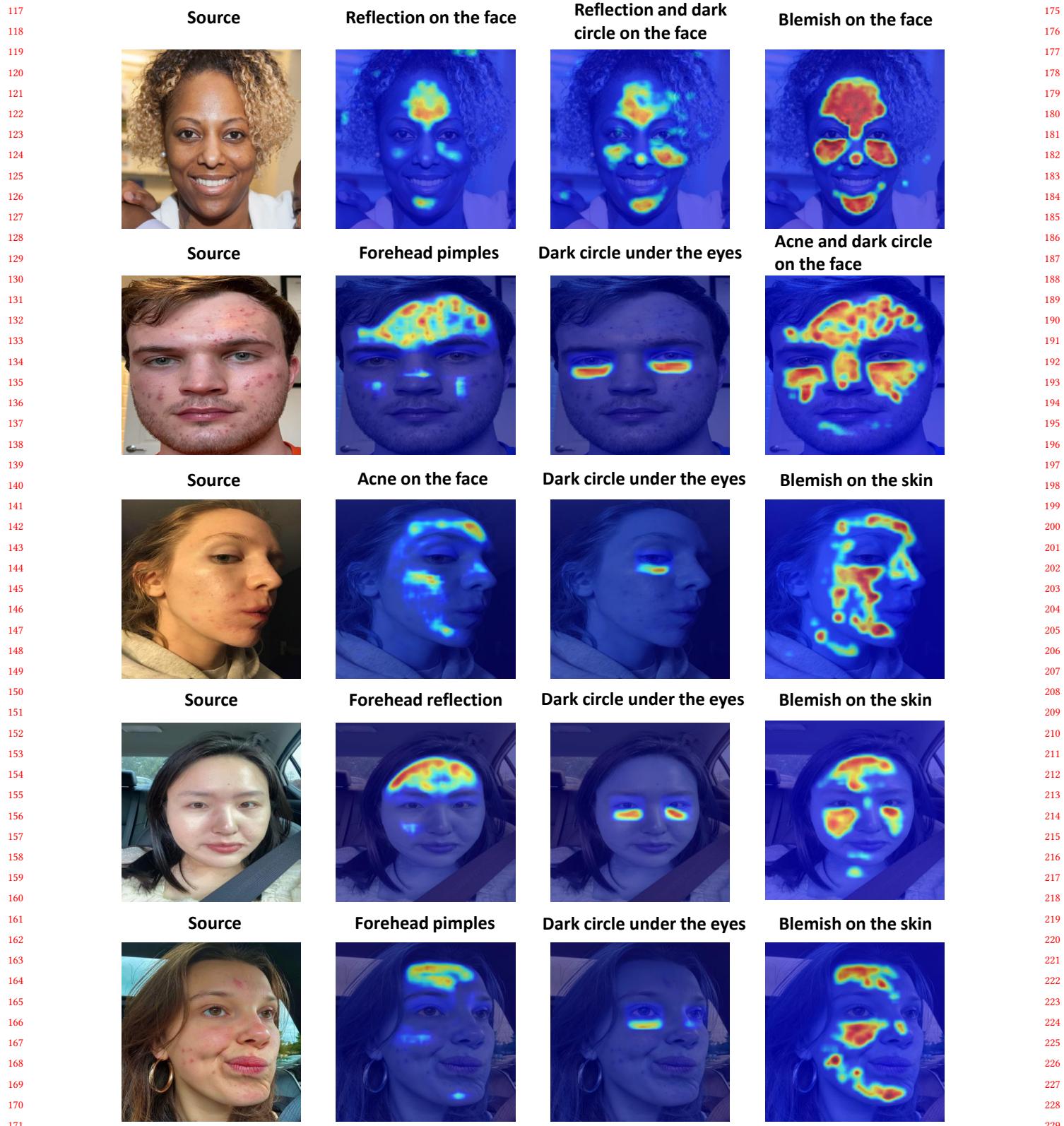
Algorithm 1 Pseudo-code of training the constituent networks in LangBRT.

```

Input: FFHQ training dataset  $X$ , pre-defined prompts  $\mathbf{pt}$  and user-defined prompts  $\mathbf{ut}$ .
Output: Clean face image  $\hat{\mathbf{y}}$ 

1: Initialize Encoder  $E$  parameterized by  $\theta_E$ , latent Transformer  $T$  parameterized by  $\theta_T$ , decoder  $D$  parameterized by  $\theta_D$ , discriminator  $S$  parameterized by  $\theta_S$ , blemish feature mapping layers  $\psi$ , weighting factor  $\gamma$  and  $\eta$ , learning rates  $\epsilon$ , and # epochs  $\Gamma$ .
2: for  $t = 1 \rightarrow \Gamma$  do
3:   for each mini-batch do
4:     Sample paired training images:  $(x, y) \in X$ .
5:     Feed the raw image  $x$  into  $E$  to obtain the feature  $F_x$ .
6:     Feed  $\{\mathbf{pt}, \mathbf{ut}\}$  into pre-trained Alpha-CLIP text encoder  $E_{Text}$  to obtain the features  $\{F_{pt}, F_{ut}\}$ , mapping  $F_{pt}$  feature with  $\psi$ , and expand the resulting embedding  $(\psi(F_{pt}) + F_{ut})$  to estimate the blemish mask  $\mathbf{m}$ , conditioned on  $F_x$ .
7:     Feed  $\mathbf{m}$  and  $F_x$  into  $T$  and  $D$  to synthesize image  $\hat{\mathbf{y}}$ .
8:     Optimize  $E$ ,  $T$  and  $D$  by using Adam:  $\{\theta_E, \theta_T, \theta_D\} \leftarrow Adam(\nabla(\mathcal{L}_{adv}^G + \mathcal{L}_{cons} - \gamma \mathcal{L}_{local} + \eta(\mathcal{L}_{detc} + \mathcal{L}_{comp})), \{\theta_E, \theta_T, \theta_D\}, \epsilon)$ 
9:     Optimize  $S$  by using Adam:  $\theta_S \leftarrow Adam(-\nabla(L_{adv}^S), \theta_S, \epsilon)$ 
10:    end for
11:  end for
12:  return  $\{\theta_E, \theta_T, \theta_D, \theta_S\}$ 

```



233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

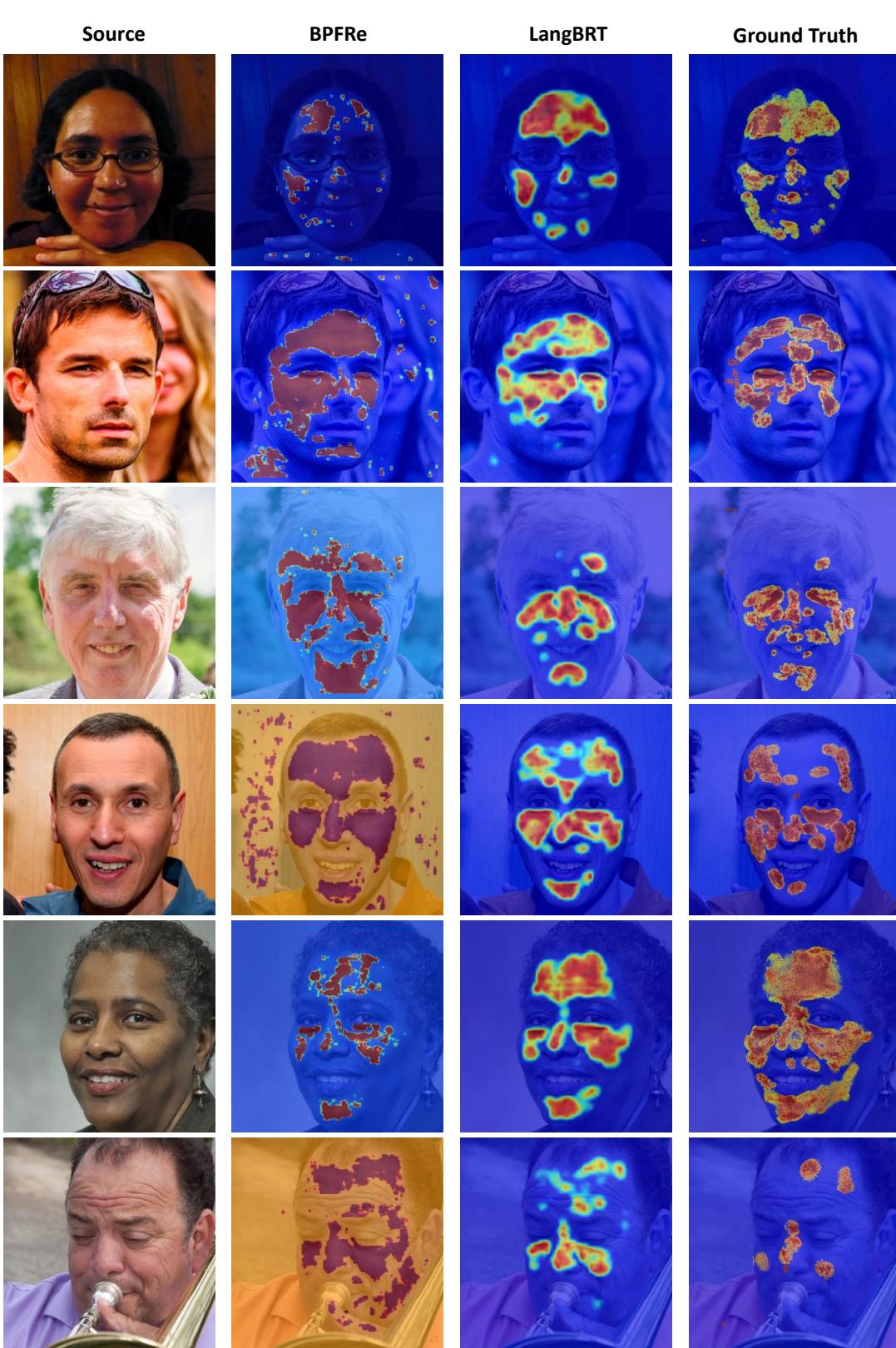
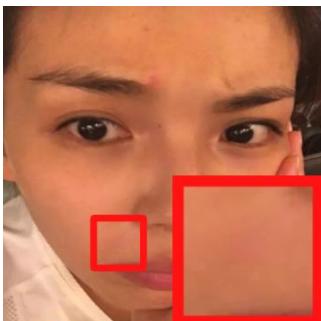


Figure 2: Visual comparison between LangBRT and BPFR in blemish detection on FFHQ images.

349

Source**ABPN****BPFRe****LangBRT**

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

1% FFHQ training data**100% FFHQ training data****1% FFHQ training data****100% FFHQ training data****1% FFHQ training data****100% FFHQ training data**

Figure 3: Representative face retouching results of LangBRT and the main competing methods on in-the-wild images.

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

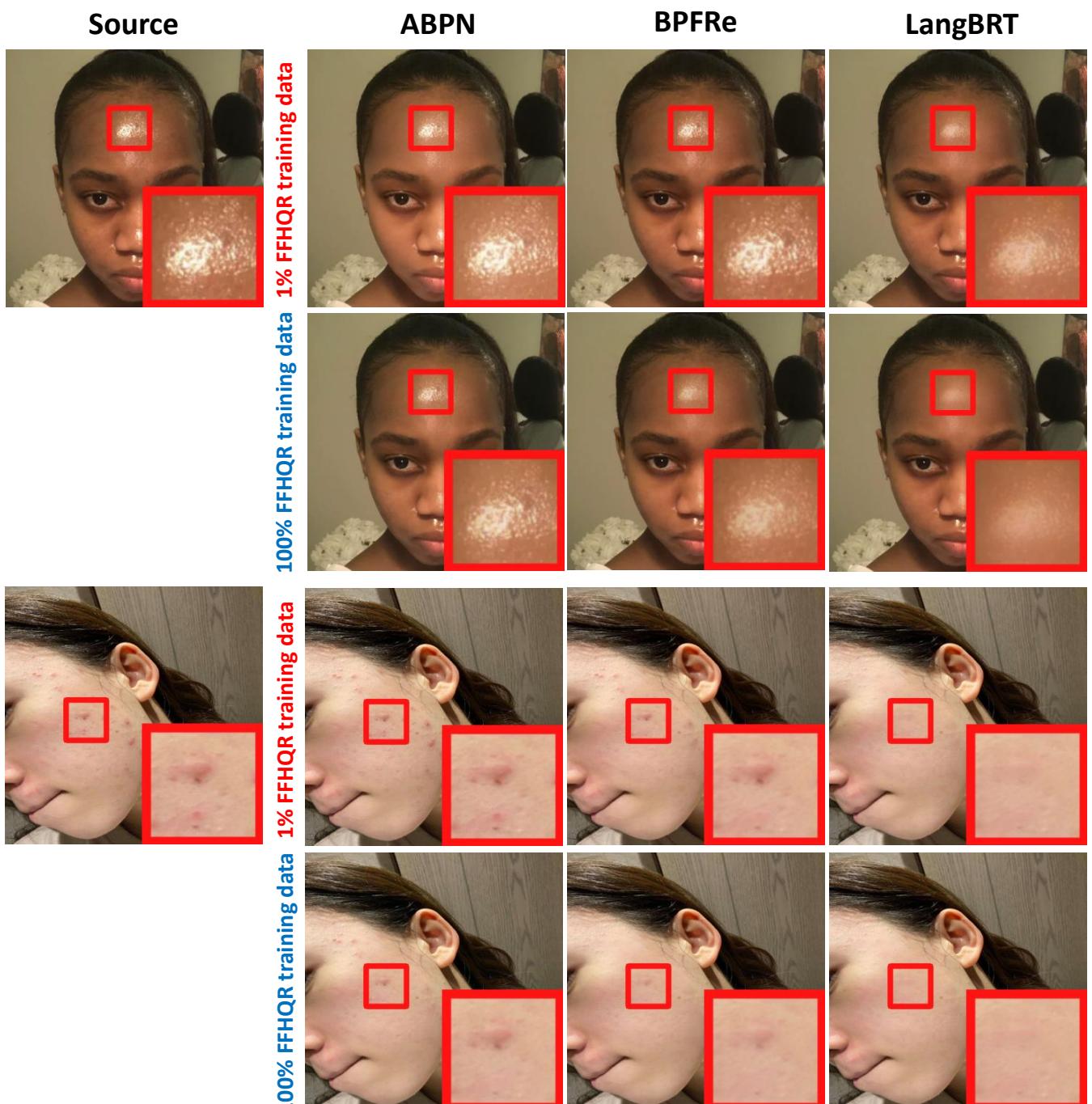


Figure 4: Representative face retouching results of LangBRT and the main competing methods on in-the-wild images.

581

582

583

584

585

586

587

588

589

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

Source**LangBRT**

Prompt: A person with (reflection on the forehead and right cheek; forehead messy hair).



612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

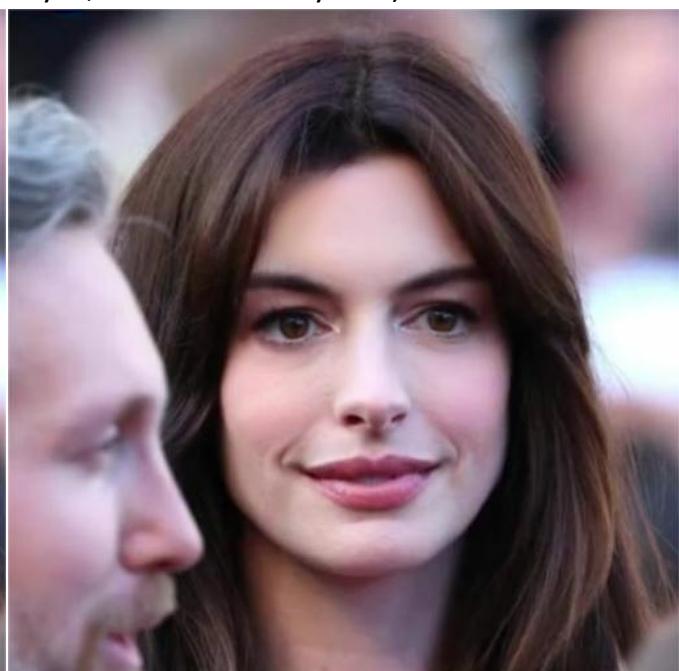


Figure 5: Reflection removal and dark circle removal results on an in-the-wild image.

697

698

699

700

701

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

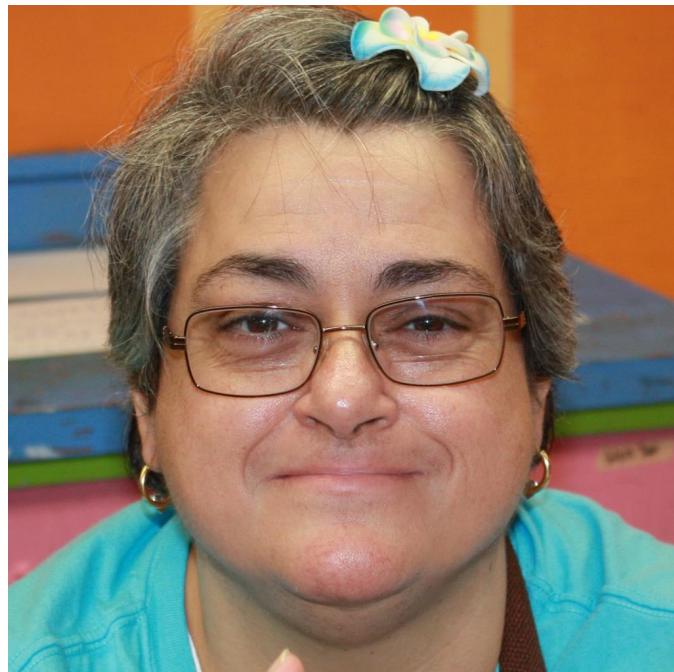
750

751

752

753

754

Source**LangBRT**

Prompt: A person with (pimples on the chin; forehead messy hair; forehead wrinkles).



Prompt: A person with (pimples on the skin; reflection on the nose).



Figure 6: Messy hair removal and Pimples removal results on an in-the-wild image.

755

756

757

758

759

760

761

762

763

764

765

766

767

768

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810

811

812

Table 1: The network architecture of the encoder E .

Layer	Activation	Input size	Output size
Conv	Leaky ReLU	$512 \times 512 \times 3$	$512 \times 512 \times 64$
Conv	Leaky ReLU	$512 \times 512 \times 64$	$256 \times 256 \times 128$
Conv	Leaky ReLU	$256 \times 256 \times 128$	$128 \times 128 \times 256$
Conv	Leaky ReLU	$128 \times 128 \times 256$	$64 \times 64 \times 512$
Conv	Leaky ReLU	$64 \times 64 \times 512$	$64 \times 64 \times 512$
Conv	Leaky ReLU	$64 \times 64 \times 512$	$64 \times 64 \times 512$

Table 2: The network architecture of the latent transformer T (TCA denotes Target-specific Cross-Attention).

Layer	Window size	Heads	Norm	MLP ratio
TCA	$6 \times 6 \times 16$	8	Layer	2.0
TCA	$6 \times 6 \times 16$	8	Layer	2.0
TCA	$6 \times 6 \times 16$	8	Layer	2.0
TCA	$6 \times 6 \times 16$	8	Layer	2.0
TCA	$6 \times 6 \times 16$	8	Layer	2.0
TCA	$6 \times 6 \times 16$	8	Layer	2.0
TCA	$6 \times 6 \times 16$	8	Layer	2.0
TCA	$6 \times 6 \times 16$	8	Layer	2.0

Table 3: The network architecture of the decoder D .

Layer	Activation	Input size	Output size
Conv	Leaky ReLU	$64 \times 64 \times 512$	$64 \times 64 \times 512$
ResBlk	Leaky ReLU	$64 \times 64 \times 512$	$64 \times 64 \times 512$
ResBlk	Leaky ReLU	$64 \times 64 \times 512$	$64 \times 64 \times 512$
ResBlk	Leaky ReLU	$64 \times 64 \times 512$	$128 \times 128 \times 256$
ResBlk	Leaky ReLU	$128 \times 128 \times 256$	$256 \times 256 \times 128$
ResBlk	Leaky ReLU	$256 \times 256 \times 128$	$512 \times 512 \times 64$
ToRGB	-	$512 \times 512 \times 64$	$512 \times 512 \times 3$

Table 4: The network architecture of the discriminator S .

Layer	Activation	Norm Layer	Input size	Output size
Conv	Leaky ReLU	Spectral Norm	$512 \times 512 \times 3$	$512 \times 512 \times 32$
Conv	Leaky ReLU	Spectral Norm	$512 \times 512 \times 32$	$256 \times 256 \times 64$
Conv	Leaky ReLU	Spectral Norm	$256 \times 256 \times 64$	$128 \times 128 \times 128$
Conv	Leaky ReLU	Spectral Norm	$128 \times 128 \times 128$	$64 \times 64 \times 256$
Conv	Leaky ReLU	Spectral Norm	$64 \times 64 \times 256$	$64 \times 64 \times 512$
Conv	Leaky ReLU	Spectral Norm	$64 \times 64 \times 512$	$64 \times 64 \times 512$