

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 054 055 056 057 058 059 060 061 062 063 064 065 066 067 068 069 070 071 072 073 074 075 076 077 078 079 080 081 082 083 084 085 086 087 088 089 090 091 092 093 094 095 096 097 098 099 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 229 230 231 232 233 234 235 236 237 238 239 239 240 241 242 243 244 245 246 247 248 249 249 250 251 252 253 254 255 256 257 258 259 259 260 261 262 263 264 265 266 267 268 269 269 270 271 272 273 274 275 276 277 278 279 279 280 281 282 283 284 285 286 287 288 289 289 290 291 292 293 294 295 296 297 298 299 299 300 301 302 303 304 305 306 307 308 309 309 310 311 312 313 314 315 316 317 318 319 319 320 321 322 323 324 325 326 327 328 329 329 330 331 332 333 334 335 336 337 338 339 339 340 341 342 343 344 345 346 347 348 349 349 350 351 352 353 354 355 356 357 358 359 359 360 361 362 363 364 365 366 367 368 369 369 370 371 372 373 374 375 376 377 378 379 379 380 381 382 383 384 385 386 387 388 389 389 390 391 392 393 394 395 396 397 398 399 399 400 401 402 403 404 405 406 407 408 409 409 410 411 412 413 414 415 416 417 418 419 419 420 421 422 423 424 425 426 427 428 429 429 430 431 432 433 434 435 436 437 438 439 439 440 441 442 443 444 445 446 447 448 449 449 450 451 452 453 454 455 456 457 458 459 459 460 461 462 463 464 465 466 467 468 469 469 470 471 472 473 474 475 476 477 478 479 479 480 481 482 483 484 485 486 487 488 489 489 490 491 492 493 494 495 496 497 498 499 499 500 501 502 503 504 505 506 507 508 509 509 510 511 512 513 514 515 516 517 518 519 519 520 521 522 523 524 525 526 527 528 529 529 530 531 532 533 534 535 536 537 538 539 539 540 541 542 543 544 545 546 547 548 549 549 550 551 552 553 554 555 556 557 558 559 559 560 561 562 563 564 565 566 567 568 569 569 570 571 572 573 574 575 576 577 578 579 579 580 581 582 583 584 585 586 587 588 589 589 590 591 592 593 594 595 596 597 598 599 599 600 601 602 603 604 605 606 607 608 609 609 610 611 612 613 614 615 616 617 618 619 619 620 621 622 623 624 625 626 627 628 629 629 630 631 632 633 634 635 636 637 638 639 639 640 641 642 643 644 645 646 647 648 649 649 650 651 652 653 654 655 656 657 658 659 659 660 661 662 663 664 665 666 667 668 669 669 670 671 672 673 674 675 676 677 678 679 679 680 681 682 683 684 685 686 687 688 689 689 690 691 692 693 694 695 696 697 698 698 699 700 701 702 703 704 705 706 707 708 709 709 710 711 712 713 714 715 716 717 718 719 719 720 721 722 723 724 725 726 727 728 729 729 730 731 732 733 734 735 736 737 738 739 739 740 741 742 743 744 745 746 747 748 749 749 750 751 752 753 754 755 756 757 758 759 759 760 761 762 763 764 765 766 767 768 769 769 770 771 772 773 774 775 776 777 778 779 779 780 781 782 783 784 785 786 787 788 789 789 790 791 792 793 794 795 796 797 798 798 799 800 801 802 803 804 805 806 807 808 809 8010 8011 8012 8013 8014 8015 8016 8017 8018 8019 8020 8021 8022 8023 8024 8025 8026 8027 8028 8029 8030 8031 8032 8033 8034 8035 8036 8037 8038 8039 8040 8041 8042 8043 8044 8045 8046 8047 8048 8049 8049 8050 8051 8052 8053 8054 8055 8056 8057 8058 8059 8059 8060 8061 8062 8063 8064 8065 8066 8067 8068 8069 8069 8070 8071 8072 8073 8074 8075 8076 8077 8078 8079 8079 8080 8081 8082 8083 8084 8085 8086 8087 8088 8089 8089 8090 8091 8092 8093 8094 8095 8096 8097 8098 8098 8099 8099 8100 8101 8102 8103 8104 8105 8106 8107 8108 8109 8109 8110 8111 8112 8113 8114 8115 8116 8117 8118 8119 8119 8120 8121 8122 8123 8124 8125 8126 8127 8128 8129 8129 8130 8131 8132 8133 8134 8135 8136 8137 8138 8139 8139 8140 8141 8142 8143 8144 8145 8146 8147 8148 8149 8149 8150 8151 8152 8153 8154 8155 8156 8157 8158 8159 8159 8160 8161 8162 8163 8164 8165 8166 8167 8168 8169 8169 8170 8171 8172 8173 8174 8175 8176 8177 8178 8179 8179 8180 8181 8182 8183 8184 8185 8186 8187 8188 8189 8189 8190 8191 8192 8193 8194 8195 8196 8197 8198 8198 8199 8199 8200 8201 8202 8203 8204 8205 8206 8207 8208 8209 8209 8210 8211 8212 8213 8214 8215 8216 8217 8218 8219 8219 8220 8221 8222 8223 8224 8225 8226 8227 8228 8229 8229 8230 8231 8232 8233 8234 8235 8236 8237 8238 8239 8239 8240 8241 8242 8243 8244 8245 8246 8247 8248 8249 8249 8250 8251 8252 8253 8254 8255 8256 8257 8258 8259 8259 8260 8261 8262 8263 8264 8265 8266 8267 8268 8269 8269 8270 8271 8272 8273 8274 8275 8276 8277 8278 8279 8279 8280 8281 8282 8283 8284 8285 8286 8287 8288 8289 8289 8290 8291 8292 8293 8294 8295 8296 8297 8298 8298 8299 8299 8300 8301 8302 8303 8304 8305 8306 8307 8308 8309 8309 8310 8311 8312 8313 8314 8315 8316 8317 8318 8319 8319 8320 8321 8322 8323 8324 8325 8326 8327 8328 8329 8329 8330 8331 8332 8333 8334 8335 8336 8337 8338 8339 8339 8340 8341 8342 8343 8344 8345 8346 8347 8348 8349 8349 8350 8351 8352 8353 8354 8355 8356 8357 8358 8359 8359 8360 8361 8362 8363 8364 8365 8366 8367 8368 8369 8369 8370 8371 8372 8373 8374 8375 8376 8377 8378 8379 8379 8380 8381 8382 8383 8384 8385 8386 8387 8388 8389 8389 8390 8391 8392 8393 8394 8395 8396 8397 8398 8398 8399 8399 8400 8401 8402 8403 8404 8405 8406 8407 8408 8409 8409 8410 8411 8412 8413 8414 8415 8416 8417 8418 8419 8419 8420 8421 8422 8423 8424 8425 8426 8427 8428 8429 8429 8430 8431 8432 8433 8434 8435 8436 8437 8438 8439 8439 8440 8441 8442 8443 8444 8445 8446 8447 8448 8449 8449 8450 8451 8452 8453 8454 8455 8456 8457 8458 8459 8459 8460 8461 8462 8463 8464 8465 8466 8467 8468 8469 8469 8470 8471 8472 8473 8474 8475 8476 8477 8478 8479 8479 8480 8481 8482 8483 8484 8485 8486 8487 8488 8489 8489 8490 8491 8492 8493 8494 8495 8496 8497 8498 8498 8499 8499 8500 8501 8502 8503 8504 8505 8506 8507 8508 8509 8509 8510 8511 8512 8513 8514 8515 8516 8517 8518 8519 8519 8520 8521 8522 8523 8524 8525 8526 8527 8528 8529 8529 8530 8531 8532 8533 8534 8535 8536 8537 8538 8539 8539 8540 8541 8542 8543 8544 8545 8546 8547 8548 8549 8549 8550 8551 8552 8553 8554 8555 8556 8557 8558 8559 8559 8560 8561 8562 8563 8564 8565 8566 8567 8568 8569 8569 8570 8571 8572 8573 8574 8575 8576 8577 8578 8579 8579 8580 8581 8582 8583 8584 8585 8586 8587 8588 8589 8589 8590 8591 8592 8593 8594 8595 8596 8597 8598 8598 8599 8599 8600 8601 8602 8603 8604 8605 8606 8607 8608 8609 8609 8610 8611 8612 8613 8614 8615 8616 8617 8618 8619 8619 8620 8621 8622 8623 8624 8625 8626 8627 8628 8629 8629 8630 8631 8632 8633 8634 8635 8636 8637 8638 8639 8639 8640 8641 8642 8643 8644 8645 8646 8647 8648 8649 8649 8650 8651 8652 8653 8654 8655 8656 8657 8658 8659 8659 8660 8661 8662 8663 8664 8665 8666 8667 8668 8669 8669 8670 8671 8672 8673 8674 8675 8676 8677 8678 8679 8679 8680 8681 8682 8683 8684 8685 8686 8687 8688 8689 8689 8690 8691 8692 8693 8694 8695 8696 8697 8698 8698 8699 8699 8700 8701 8702 8703 8704 8705 8706 8707 8708 8709 8709 8710 8711 8712 8713 8714 8715 8716 8717 8718 8719 8719 8720 8721 8722 8723 8724 8725 8726 8727 8728 8729 8729 8730 8731 8732 8733 8734 8735 8736 8737 8738 8739 8739 8740 8741 8742 8743 8744 8745 8746 8747 8748 8749 8749 8750 8751 8752 8753 8754 8755 8756 8757 8758 8759 8759 8760 8761 8762 8763 8764 8765 8766 8767 8768 8769 8769 8770 8771 8772 8773 8774 8775 8776 8777 8778 8779 8779 8780 8781 8782 8783 8784 8785 8786 8787 8788 8789 8789 8790 8791 8792 8793 8794 8795 8796 8797 8798 8798 8799 8799 8800 8801 8802 8803 8804 8805 8806 8807 8808 8809 8809 8810 8811 8812 8813 8814 8815 8816 8817 8818 8819 8819 8820 8821 8822 8823 8824 8825 8826 8827 8828 8829 8829 8830 8831 8832 8833 8834 8835 8836 8837 8838 8839 8839 8840 8841 8842 8843 8844 8845 8846 8847 8848 8849 8849 8850 8851 8852 8853 8854 8855 8856 8857 8858 8859 8859 8860 8861 8862 8863 8864 8865 8866 8867 8868 8869 8869 8870 8871 8872 8873 8874 8875 8876 8877 8878 8879 8879 8880 8881 8882 8883 8884 8885 8886 8887 8888 8889 8889 8890 8891 8892 8893 8894 8895 8896 8897 8898 8898 8899 8899 8900 8901 8902 8903 8904 8905 8906 8907 8908 8909 8909 8910 8911 8912 8913 8914 8915 8916 8917 8918 8919 8919 8920 8921 8922 8923 8924 8925 8926 8927 8928 8929 8929 8930 8931 8932 8933 8934 8935 8936 8937 8938 8939 8939 8940 8941 8942 8943 8944 8945 8946 8947 8948 8949 8949 8950 8951 8952 8953 8954 8955 8956 8957 8958 8959 8959 8960 8961 8962 8963 8964 8965 8966 8967 8968 8969 8969 8970 8971 8972 8973 8974 8975 8976 8977 8978 8979 8979 8980 8981 8982 8983 8984 8985 8986 8987 8988 8989 8989 8990 8991 8992 8993 8994 8995 8996 8997 8998 8998 8999 8999 9000 9001 9002 9003 9004 9005 9006 9007 9008 9009 90010 90011 90012 90013 90014 90015 90016 90017 90018 90019 90020 90021 90022 90023 90024 90025 90026 90027 90028 90029 90030 90031 90032 90033 90034 90035 90036 90037 90038 90039 90039 90040 90041 90042 90043 90044 90045 90046 90047 90048 90049 90049 90050 90051 90052 90053 90054 90055 90056 90057 90058 90059 90059 90060 90061 90062 90063 90064 90065 90066 90067 90068 90069 90069 90070 90071 90072 90073 90074 90075 90076 90077 90078 90079 90079 90080 90081 90082 90083 90084 90085 90086 90087 90088 90089 90089 90090 90091 90092 90093 90094 90095 90096 90097 90098 90098 90099 90099 90100 90101 90102 90103 90104 90105 90106 90107 90108 90109 90109 90110 90111 90112 90113 90114 90115 90116 90117 90118 90119 90119 90120 90121 90122 90123 90124 90125 90126 90127 90128 90129 90129 90130 90131 90132 90133 90134 90135 90136 90137 90138 90139 90139 90140 90141 90142 90143 90144 90145 90146 90147 90148 90149 90149 90150 90151 90152 90153 90154 90155 90156 90157 90158 90159 90159 90160 90161 90162 90163 90164 90165 90166 90167 90168 90169 90169 90170 90171 90172 90173 90174 90175 90176 90177 90178 90179 90179 90180 90181 90182 90183 90184 90185 90186 90187 90188 90189 90189



Fig. 1. We present a model which aims to understand different aspects of the generalized attention prediction problem which are exemplified above. In (a) subjects are looking at a salient object in the scene, in (b) the subject is looking somewhere outside of the scene and in (c) the subject is looking at or around the camera. Our model sets out to predict the 3D gaze vector of subjects in each of these images as well as the location of the gaze fixation in the image, if it exists. If not, the subject is looking somewhere outside of the scene and the model should predict this case as well.

work of Recasens et al. [4], which tackles this problem by obtaining human annotations of subject gaze targets, leveraging the finding from [8–10] which indicate that annotators very often agree on which object is salient in the scene. Their approach, however, was not designed to handle cases (b) and (c) since the dataset annotation process forces human annotators to label a point in the image as the fixation location. In other words the dataset does not distinguish between subjects looking at a point inside of the image or looking somewhere outside of the image. A purely saliency based approach would also fail: notice that there are salient objects in (b), an American flag, and (c), a mug, which can confound such an approach.

Figure 1-(c) resembles the case of Kraftka et al. [3] and has been tackled in other works as well thanks to the creation of datasets in which subjects look at objects in screens and where all elements are calibrated such that the direction of the gaze can be obtained in a reliable manner [11]. Figure 1-(b) presents the hardest case - it is very difficult to annotate similar images since 3D gaze direction is required. For example one way to do this is to map the 3D scene as well as capture the 3D location of the observed point in order to obtain the gaze direction. Naturally, we have not found a dataset which addresses this aspect of the generalized attention prediction problem.

Indeed, obtaining a system that can deal with these different cases is a hard problem. This paper sets out to solve this by presenting a novel generalized visual attention estimation method which jointly learns a subject-dependent saliency map and the 3D gaze vector represented by yaw and pitch. This allows us to estimate the final fixation likelihood map.

We design the method to produce the output in such a way that the fixation likelihood map becomes close to zero when the subject is looking outside the frame as in cases (b) and (c). When the subject is looking at a point inside of

the scene, similar to case (a), then the fixation likelihood map predicts where the subject is likely to attend in the image. The model simultaneously estimates the 3D gaze angle to provide a complete picture of the subject’s attention and gaze. As a result, our approach produces interpretable results spanning all three cases in Figure 1.

Our Contribution. The major contribution of this work is the attention estimation method which addresses the **generalized visual attention prediction problem** and which works across most natural scenarios. To effectively train our model we exploit three public datasets that have been originally collected for different tasks. Specifically, we use the EYEDIAP dataset [11] to learn precise gaze angle representation, a modified version of GazeFollow dataset [4] to learn gaze-relevant scene saliency representation, and the SynHead dataset [12] to complement the first two datasets as it includes large face pose variations and subject attention outside of the image.

As a result of our multi-task learning approach, our model achieves state-of-the-art results on the GazeFollow [4] task, which consists in identifying the location of the scene the subject is looking at. Our model also competes with state-of-the-art models on the 3D gaze estimation task on the EYEDIAP dataset [11]. Most importantly, we evaluate our full model on a new challenging task that automatically quantifies dense visual attention in naturalistic social interactions. We report our results on the Multimodal Dyadic Behavior (MMDB) dataset [13], a dataset of video recordings of the social and communicative behavior of toddlers. This dataset has frame-level annotations of subject’s visual targets among many other nonverbal behaviors. We are the first to report attention estimation results on this dataset. We compare our results to a variety of baseline tests that are outperformed by our method.

2 Related work

Gaze Estimation: Gaze estimation aims to predict the gaze of a human subject. Our work is related to third person gaze estimation and tracking methods which seek to estimate either the three dimensional direction of the gaze or the fixation point of the gaze on a screen. Krafska et al. [3] predict the coordinates of the gaze of a person on a smartphone device and present a dataset which addresses this problem. Mora et al. [11] present EYEDIAP, a dataset designed for the evaluation of gaze estimation task and collected in a controlled lab environment. They implement a RGB-D method which predicts the 3D vector of the gaze of the subject. There exist datasets which address the same task such as MPIIGaze [2] as well as synthetic datasets of eye images for gaze estimation [14, 15]. In addition to predicting the 3D gaze vector our work predicts a fixation likelihood map of the scene as well as whether the person is looking at a location inside or outside of the image.

Visual Saliency: The objective of visual saliency prediction is to estimate locations in an image which attract the attention of humans looking *at* the image. Since the seminal work of Itti et al. [16] visual saliency prediction has been

extensively studied. Recently deep learning methods have shown superior performance on this task due to their ability to learn features and to incorporate both local and global context into the prediction [17–19]. Our work in generalized visual attention prediction is influenced by the task of visual saliency since people tend to look at salient objects inside a scene, yet it is distinct because we consider cases where the subject is not looking at any object in the scene. A method driven mainly by saliency detection would not be able to succeed in these scenarios. Furthermore, a person actively involved in a scenario will often be looking at a different location than a human who is free-viewing an image.

Gaze Following: The paper by Recasens et al. [4] presents a new computer vision problem which deeply inspired our work. The problem can be described as follows: given a single image containing one or more people, predict the location that each person in the scene is looking at. Their work presents a novel dataset which contains human annotations of where the subjects are looking at inside of the scene. Our work differs from this previous work in that we consider cases where the subjects are looking outside of the scene and we predict these cases. In addition to predicting a fixation likelihood map for the image we predict the 3D gaze vector of the subject. Gorji and Clark [20] study a problem in the intersection of visual saliency and gaze following which consist in incorporating signals from other regions of the image which push attention to a certain part of the image. For example when subjects in an image look at a salient object in the image this amplifies the apparent saliency of the object. Again, our problem differs in that we do not predict visual saliency but we predict gaze fixation and gaze direction of subjects inside of the scene.

Attention Modeling: People have proposed different methods for measuring third-person visual attention from environment-mounted camera. By assuming body or head orientation is a good proxy to approximate visual orientation, [21] projects attention on the street by tracking pedestrians in 3D, [22, 23] model the focus of attention in crowded social scene, and [24] predicts the object in the scene that a person is interacting with which is usually indicated by hand manipulation or pointing. Our work is certainly related, although it differs because we explicitly consider the gaze of the subject.

3 Method

Figure 2 is an overview of our deep neural network model and its input and output. The model takes three inputs: the whole image, a crop of the subject’s face, and the location of her face. Given the input, the model estimates 1. the subject’s gaze angle in terms of yaw and pitch degrees (“where” component of visual attention), 2. the subject-dependent saliency in terms of a heatmap (“what” component of visual attention), and 3. how likely the subject is fixating at the estimated gaze target in the scene (overall “strength” of visual attention).

The model has two fully-convolutional pathways, one connected to the whole image (Figure 2-a) and the other connected to the face image (Figure 2-b). The reasoning behind having two separate pathways is inspired by the way humans

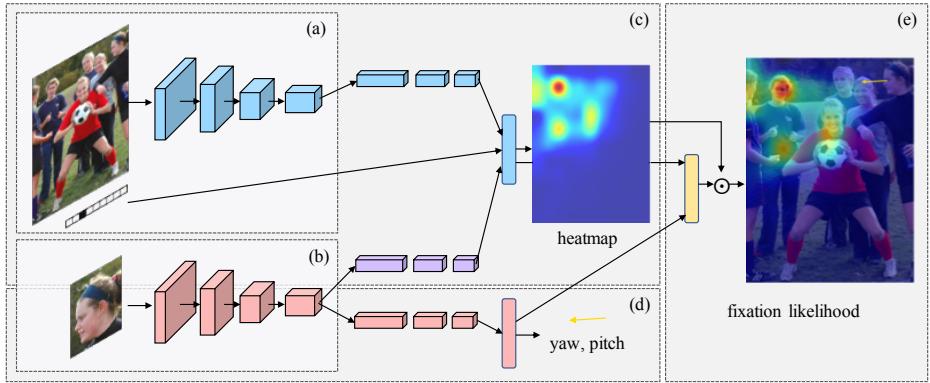


Fig. 2. Overview of our approach. Full scene image, a person's face location whose visual attention we want to predict, and the corresponding close-up face image is provided as input. Scene and face images go through separate convolutional layers in such a way that (a) (b) and (c) contribute to person-centric saliency, and (b) and (d) contribute to gaze angle prediction. In the very last layer, the final feature vectors for these two tasks are combined to estimate how likely the person is actually fixating at a gaze target within the observable scene.

infer one's visual attention, as first suggested by [4]. For example, when we interpret a person's attention from an image, we generally guess where the person's gaze is directed toward and consider if there is any salient part in the scene that lies along the direction. Based on this hypothesis, [4] connect two independent conv pathway together to learn the heatmap (Figure 2-c). We take this idea further and extend their work by explicitly training for gaze angle (Figure 2-d) with the convolutional pathway that is connected to the face image, in a multi-task learning framework. Adding the gaze angle output as an auxiliary task has several advantages including the additional supervisory signal that we can devise based on the relationship between gaze heatmap and angle, which pushes performance in heatmap estimation even further.

Lastly, we define the likelihood of fixation: a single-valued measure of how likely the subject is looking at the estimated target region inside the frame. It is modeled by a fully connected layer (Figure 2-e). Using this last output, the model can now produce a much more complete estimation of a person's visual attention. Think of the case of Figure 1-b or c, where the person is looking outside the image frame. In such cases, we want the heatmap to be as close to zero as possible since the person is not attending to any point inside of the image. By training this last layer to produce higher value for when it is more certain that the heatmap region is attended to and lower value otherwise, the value can be applied to the heatmap with an operator \odot which can be a weighting operator or a gating operator depending on application.

Since there exists no single dataset that covers all various gaze and scene compositions that we consider in this paper (e.g., looking outside the frame,

looking around the camera, clear fixation on an object, etc), we adopt a cross domain learning approach where the model learns partial information relevant to each task from different datasets. Depending on what supervisory signal is available in a given batch of training data, the model selectively updates its corresponding branches.

We describe the model architecture in more detail in subsection 3.1. We elaborate on the loss function in 3.2 and talk about the datasets and training procedures in 3.3.

3.1 Model

The inputs given to the model are the entire image, the subject’s cropped face and the location of the subject’s face whose attention we want to estimate. The two images are resized to 227×227 so that the face can be observed in higher resolution by the network. Face position is available in terms of the (x, y) full image coordinates. These coordinates are quantized into a 13×13 grid and then flattened to a 169 dimensional 1-hot vector.

The model consists of two convolutional pathways: a face pathway (Figure 2-d) and a scene pathway (Figure 2-c). ResNet 50 [25] is used as a backbone network for the convolutional pathways (Figure 2-a,b). Specifically we use all convolutional layers of ResNet50 for each of our convolutional pathways. After each ResNet50 block we add three conv layers (1×1 , followed by 3×3 , followed by 1×1) with ReLu and Batchnorm - with stride 1 and no padding. The blue conv layers represented in (c) have filter depth of 512, 128 and 1 respectively. The purple and red conv layers after the face pathway (represented in (c) and (d)) have filter depth of 512, 128 and 16. These conv layers serve to reduce the dimensionality of the features extracted by the ResNet50 backbone networks.

In the face pathway, the feature vector computed with the face input image goes through a fully connected layer to predict the gaze angle represented using yaw and pitch intrinsic Euler angles. In the scene pathway, the feature vectors extracted from the whole image as well as from the face image are concatenated with the face position input vector to learn the person-centric heatmap. Similarly to face position, the ground truth used for learning the heatmap is available as a gaze target position in (x, y) coordinates which is quantized into a 10×10 grid and then flattened to a 100-dimensional one-hot vector for training.

Lastly, the input vectors to the last layer of each pathway are concatenated and go into the final fully connected layer to estimate the “strength” of the fixation ie. how likely it is that the person is actually fixating at a gaze target within the observable scene. The training label for this value is equal to 1 for a fixation inside of the image and 0 when the subject is looking outside of the scene. We also explore alternative model architectures and restrict our training to a subset of the three datasets. Experiments are reported in Section 4.4.

3.2 Loss

As our model predicts gaze angle, saliency map and the fixation likelihood, we need to apply appropriate loss functions for each task. For the angle regression

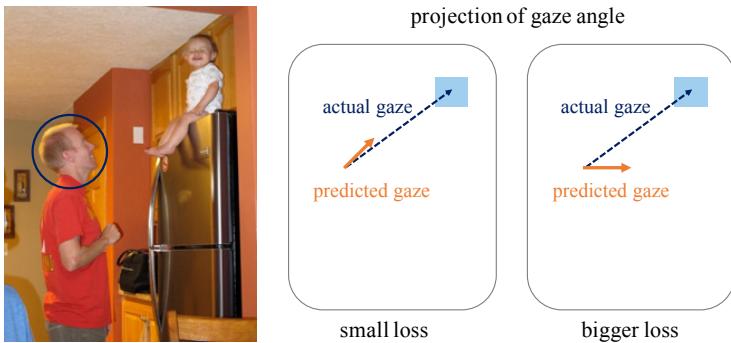


Fig. 3. Our project and compare loss is illustrated here. If the estimated angle is close to the actual one, the projected gaze angle on the image should also be close to the vector connecting the head position to the gaze target.

task we use an **L1 loss**, and for the other two tasks we use a **cross entropy loss**. Moreover, we recognize that the gaze angle and fixation target predictions are closely related. Based on their relationship additional constraints can be imposed to augment the training loss signal. Namely, when the subject is looking at a target, the actual gaze is a ray coming from the subject's head to the gaze target. This ray can be projected onto the image. It becomes a 2D vector coming from the subject's head to the target exemplified by the blue vector in Figure 3. If the estimated angle is close to the actual one, the projected gaze angle on the image (orange vector in Figure 3) should also be close to the blue vector. The closeness between the two vectors is measured using a cosine distance which supplements our learning procedure. We call this the **project and compare loss**.

3.3 Cross-Domain Datasets and Training Procedure



Fig. 4. Examples of datasets used to train our model. Left two: SynHead, middle two: EYEDIAP, right two: GazeFollow.

The largest challenge in training our model is the availability of training examples. Although there are a couple of existing datasets that are suitable for training certain parts of our network, no unique dataset contains all of the information that we need to train the full model. Therefore, we leverage three different

315 types of datasets, namely, GazeFollow [4], EYEDIAP [11], and SynHead [12]. We
316 selectively train different sub-parts of our network at a time depending on the
317 available supervisory information within a training batch. See Figure 4 to see
318 sample images from each dataset.

319 GazeFollow [4] is a real-world dataset with manual annotations of the location
320 of where people in images are looking. The images used in this dataset come from
321 other major datasets such as MS COCO [26] and PASCAL [27]. As a result, the
322 images cover a wide range of scenes, people, and gaze directions. However, the
323 actual 3D gaze angles are not available. Furthermore, images where subjects are
324 looking outside of the image frame are not distinguished and all images have a
325 fixation annotation inside of the frame. Although it is mentioned in [4] that if the
326 annotators indicated that the person was looking outside the image, the image
327 would be discarded, we notice that there are a considerable number of images in
328 which persons are looking outside of the frame. Therefore, we added additional
329 annotations to this dataset in the form of a binary indicator label for “looking
330 inside” or “looking outside” for every image. In total, we identified 14,564 images
331 correspond to the “looking outside” case which is approximately 11.6% of the
332 total training samples. We publicly release our additional annotations along with
333 this paper.

334 EYEDIAP [11] dataset is designed for the evaluation of the gaze estimation
335 task. It has videos of 16 different subjects with full face and background visible
336 in a laboratory environment. Each subject was asked to look at a specific target
337 point on a monitor screen and the 3D gaze angle was annotated by leveraging
338 camera calibration and face depth measurement from depth camera. This
339 dataset contains precise 3D gaze angles for frames where the person is fixating
340 the target point. The dataset also contains video of the subjects looking at a 3D
341 ball target instead of 2D screen target point, but we exclude these ball sessions
342 from our experiments in order to conduct a fairer comparison with prior work.
343 We randomly hold out four subjects for test and use the rest of the sessions
344 for training. Since subjects were looking at a screen, all of the frames can be
345 considered as looking outside the image. However, since the dataset has been
346 collected in a controlled settings the background is mostly white and clear and
347 there is not a lot of variety in lighting or pose. Also, measured gaze angles range
348 between -40° to 40° which is rather limited.

349 NVIDIA SynHead [12] is a synthetic dataset created for the head pose es-
350 timation task. The dataset contains 510,960 frames of 70 head motion tracks
351 rendered using 10 individual head models. The gaze of the head is fixed and
352 aligned with the head pose, thus we use the labeled 3D head pose as the gaze
353 angle ground truth. One of the advantages of a synthetic dataset is the ability to
354 insert different images in the background. We randomly generated 15% from the
355 total frames augmented with provided natural scene backgrounds and regard all
356 as “looking outside” examples. The main reason we include SynHead in training
357 is because it complements the EYEDIAP dataset, as the angle ranges are larger,
358 between -90° and 90° , and it can include more diverse backgrounds. Since head
359 pose estimation is not a focus of this paper we do not set aside a test set and use
SynHead entirely for training. Dataset details are also summarized in Table 1.

Table 1. Datasets used in our experiments and the number of samples in the training and testing split, as well as the percentage of each split containing people looking in/out.

Dataset	Training set		Test set	
		in vs out		in vs out
GazeFollow [4]	125,557	88.4% vs 11.6%	4,782	100% vs 0%
EYEDIAP [11]	72,613	0% vs 100%	18,153	0% vs 100%
SynHead [12]	75,400	0% vs 100%	-	-
MMDB [13]	-	-	4,965	41.4% vs 58.6%

Training Procedure. Since each dataset is relevant only to certain sub-tasks, we only update the relevant parts of the network based on which dataset the training sample is from, while freezing other irrelevant layers during back-propagation. Specifically, when learning gaze angle estimation, we only update the angle pathway (b) and (d) in Figure 2, when learning saliency we update the scene pathway (a), (b) and (c) while freezing all other layers. Similarly, when training fixation likelihood we only update the layer (e) in Figure 2. We found that this selective backpropagation scheme is critical in achieving good performance.

In every batch, we draw random samples from all of the datasets shuffled together and perform three separate backpropgataions for the three outputs as just described. In the beginning, both convolutional pathways were initialized using a ResNet50 model pretrained on the ImageNet classification task [28]. We use the Adam optimization algorithm with a learning rate of $2.5e-4$ and a batch size of 36. Training usually converges within 12 epochs. All of our implementation and experiments are done in PyTorch [29].

4 Evaluation

In this section we evaluate our results by comparing each output with a number of existing methods and baselines. We first evaluate the person-dependent saliency map in 4.1, gaze angle estimation in 4.2 and general attention estimation in 4.3. Lastly, we evaluate our method by changing model architectures and training dataset in 4.4.

4.1 Person-Dependent Saliency Prediction

We evaluate the performance of saliency map estimation using the suggested test split of the GazeFollow dataset. The test split contains all “looking inside” cases and each test image has multiple gaze target annotations. Following the same evaluation method by [4], we compute the Area Under Curve (AUC) score of the Receiver Operating Characteristic (ROC) curve in which the ground truth



Fig. 5. Qualitative results of our model’s gaze-saliency prediction on the GazeFollow dataset. Input image is given on the 1st and 3rd row, the output heatmap and estimated gaze is overlaid below.

Table 2. Gaze-saliency evaluation on the GazeFollow test set

Method	AUC	L2 Distance	Min Distance	Angle
Random	0.504	0.484	0.391	69°
Center	0.633	0.313	0.230	49°
Judd [30]	0.711	0.337	0.250	54°
GazeFollow [4]	0.878	0.190	0.113	24°
Our	0.896	0.187	0.112	23°

target positions are the true labels and heatmap value on corresponding positions are prediction confidence score. Our method achieves a score of 0.896 achieving state-of-the-art performance. Along with AUC we also report results in L2, min distance and angle metric. Please refer to [4] for details about the metric. The numbers are summarized in Table 2 along with a number of baselines reported in [4]. Qualitative results are presented in Figure 5.

4.2 Gaze Angle Prediction

We report the 3D gaze estimation accuracy based on the yaw and pitch output of our model on the chosen EYEDIAP test split. Table 3 shows the angular errors in which we achieve less than 0.5 degrees of difference to the state-of-the-art appearance-based gaze estimation method. It is worth noticing that the middle two values come from [31] which are computed by five-fold cross validation with the entire EYEDIAP dataset whereas our method is evaluated on a single train/test split. Although we did not choose to perform full cross validation, we

Table 3. Gaze angle evaluation on EYEDIAP

Method	Angular Error (degree)
Wood [14]	11.3°
iTracker [3]	8.3°
Zhang [31]	6.0°
Our	6.4°

conclude that it reaches reasonable accuracy on the benchmark. Note also that our method is trained on multiple tasks whereas all other methods are trained solely on the gaze angle prediction task.

4.3 Generalized Attention Prediction During Naturalistic Social Interactions

The primary inspiration for our work stems from the need for the ability to quantify various types of visual attention behavior, which is one of the most important nonverbal social cues used in our daily life. Moreover, this is of particular interest among researchers who study child development since gaze behavior of young children is closely related to their social development and developmental disorders such as Autism [32]. The MMDB dataset is one of the largest datasets that contains children’s social and communicative behaviors, collected in order to facilitate data-driven analysis of child behavior based on video. The dataset contains a wide range of nonverbal behavior such as hand gestures, smile, and gaze. It has frame-level human annotations of each behavior. As for gaze, each frame is annotated when the child is looking at a ball, book or the examiner. This is done by human annotation based on multiple views, therefore the child’s gaze target can be visible or not depending on the viewpoint. Since the annotation does not indicate in which view the gaze target is visible, we added additional annotation ourselves and identified if the target is visible in a child-facing camera view to construct labels for the general attention estimation problem. We publicly release this annotation text file along with our paper.

We evaluate our method on the generalized attention prediction task. We design a gaze target grid classification task, where each test image is divided into $N \times N$ grids. If the subject is looking inside of the image then the grid square which contains the gaze target is assigned a label of 1 while others are assigned labels of 0. If the subject is looking somewhere outside of the frame then all grid squares are assigned the 0 label. Using our method’s fixation likelihood map we predict the positive gaze grid square. We test the GazeFollow model [4] which is the closest work to our method in terms of having the ability to predict gaze target location. One of its limitations is the inability to correctly predict the “outside” case, where the subject is looking outside of the frame. As a result, our method achieves much higher precision in addition to increased recall as shown in Table 4.

495 Additionally, we constructed various baseline tests consisting of a classifier
 496 based on a subset of features constructed for saliency, gaze and head pose. Specifi-
 497 cally, we tested with SVM and Random Forest using a subset of {[4], [31], [33]}
 498 as features. In other words, each classifier has been trained for detection of look-
 499 ing inside with the training set described in Table 1, using one or more of the
 500 three methods’ output, and tested on the MMDB images. We report the results
 501 in Table 5. Note that the MMDB dataset was not used for training across all
 502 methods including ours.

503
 504 **Table 4.** Evaluation on MMDB - gaze target grid classification
 505

506 Grid Size	507 Method	508 Precision	509 Recall
508 2x2	GazeFollow [4]	0.344	0.715
	Our	0.744	0.851
510 5x5	GazeFollow [4]	0.210	0.437
	Our	0.614	0.683

512
 513 **Table 5.** Evaluation of fixation likelihood on MMDB
 514
 515

516 Method	517 Average Precision
SVM with GazeFollow [4]	0.311
SVM with GazeFollow [4]+gaze [31]	0.531
SVM with GazeFollow [4]+headpose [33]	0.620
SVM with gaze [31]+headpose [33]	0.405
SVM with GazeFollow [4]+gaze [31]+headpose [33]	0.624
Random Forest with GazeFollow [4]	0.707
Random Forest with GazeFollow [4]+gaze [31]	0.727
Random Forest with GazeFollow [4]+headpose [33]	0.785
Random Forest with gaze [31]+headpose [33]	0.512
Random Forest with GazeFollow [4]+gaze [31]+headpose [33]	0.773
Our	0.902

533 4.4 Alternative Model and Diagnostics

534 Finally, we run additional experiments to study how the performance of our
 535 model is affected by different training datasets and architectural choice. As
 536 shown in Table 6, omitting EYEDIAP or SynHead training dataset did not
 537 have much impact on the heatmap estimation whereas changing model archi-
 538 tecture considerably affected the scores. For example using a single ResNet50
 539

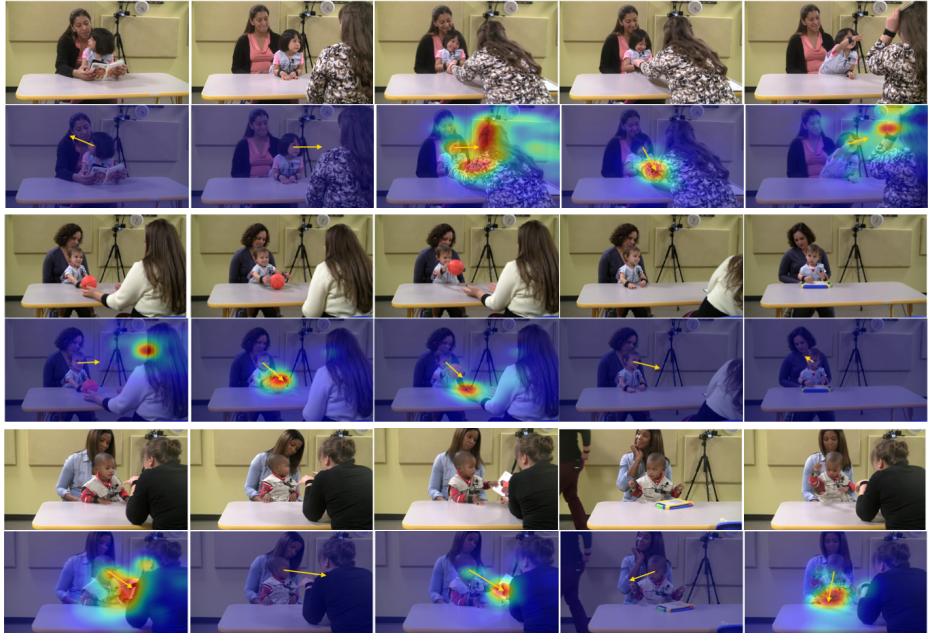


Fig. 6. Example result of our method on the MMDB dataset. The dataset contains various types of gaze behavior including fixations on a target both within and out of frame. Our method produces low heatmap when the fixation target is outside and high heatmap when the target becomes clear.

pathway which pools facial features using ROI-pooling shows significantly degraded performance which supports our decision to use a scene pathway as well as a face pathway. Qualitatively, we were able to observe that, even though our method is designed to measure fixation outside, it can make mistakes when the target is within the frame but occluded by other object. Also, when the subject is closer to the camera than some salient object in the background, the method sometimes estimates those as fixation candidate due to the lack of scene depth understanding. Examples are illustrated in Figure 7.

5 Conclusion

In this paper we present the new challenging problem of generalized visual attention prediction which encapsulates several constrained attention prediction and gaze estimation problems that have been the focus of many previous works. We propose a multi-task learning approach and neural architecture leveraging three different datasets which tackles this problem and works across multiple naturalistic social scenarios. In order to train our architecture we have supplemented these datasets with new annotations and we release these annotations to the public. Our model achieves state-of-the-art performance on the single-task

Table 6. Additional model evaluation and diagnostics

Method	GazeFollow test	
	AUC	L2 Distance
No EYEDIAP	0.887	0.197
No SynHead	0.895	0.191
No EYEDIAP and SynHead	0.891	0.194
No project-and-compare loss	0.895	0.189
Map resolution 15x15	0.778	0.194
ROI-pooling	0.700	0.325
Our final	0.896	0.187

**Fig. 7.** Challenging cases due to occlusion and the lack of depth understanding.

gaze-saliency prediction and competes with state-of-the-art methods on gaze estimation benchmarks while achieving promising performance on the generalized attention prediction problem on the MMDB dataset. Future work in this area can lead to breakthroughs in attention prediction applications which are valuable in numerous scientific and commercial areas. A suggested first step would be to improve existing datasets with additional annotations or collect datasets tailored for this problem.

630 References

- 631 1. Land, M., Tatler, B.: Looking and acting: vision and eye movements in natural
632 behaviour. Oxford University Press (2009)
- 633 2. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation
634 in the wild. In: Proc. of the IEEE Conference on Computer Vision and Pattern
635 Recognition (CVPR). (June 2015) 4511–4520
- 636 3. Krafska, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W.,
637 Torralba, A.: Eye tracking for everyone. In: IEEE Conference on Computer Vision
638 and Pattern Recognition (CVPR). (2016)
- 639 4. Recasens*, A., Khosla*, A., Vondrick, C., Torralba, A.: Where are they looking?
640 In: Advances in Neural Information Processing Systems (NIPS). (2015) * indicates
641 equal contribution.
- 642 5. Chong, E., Chanda, K., Ye, Z., Southerland, A., Ruiz, N., Jones, R.M., Rozga, A.,
643 Rehg, J.M.: Detecting gaze towards eyes in natural social interactions and its use
644 in child assessment. Proceedings of the ACM on Interactive, Mobile, Wearable and
645 Ubiquitous Technologies 1(3) (2017) 43
- 646 6. Zhang, X., Sugano, Y., Bulling, A.: Everyday eye contact detection using unsupervised
647 gaze target discovery. In: 30th Annual Symposium on User Interface Software
648 and Technology, ACM (2017)
- 649 7. Recasens, A., Vondrick, C., Khosla, A., Torralba, A.: Following gaze in video. In:
650 The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
- 651 8. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object
652 segmentation. In: Proceedings of the IEEE Conference on Computer Vision and
653 Pattern Recognition. (2014) 280–287
- 654 9. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: A benchmark.
655 IEEE Transactions on Image Processing 24(12) (2015) 5706–5722
- 656 10. Borji, A., Sihite, D.N., Itti, L.: What stands out in a scene? a study of human
657 explicit saliency judgment. Vision research 91 (2013) 62–77
- 658 11. Funes Mora, K.A., Monay, F., Odobez, J.M.: Eyediap: A database for the develop-
659 ment and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In:
660 Proceedings of the ACM Symposium on Eye Tracking Research and Applications,
661 ACM (March 2014)
- 662 12. Gu, J., Yang, X., De Mello, S., Kautz, J.: Dynamic facial analysis: From bayesian
663 filtering to recurrent neural network. In: The IEEE Conference on Computer Vision
664 and Pattern Recognition (CVPR). (July 2017)
- 665 13. Rehg, J., Abowd, G., Rozga, A., Romero, M., Clements, M., Sclaroff, S., Essa,
666 I., Ousley, O., Li, Y., Kim, C., et al.: Decoding children’s social behavior. In:
667 Proceedings of the IEEE conference on computer vision and pattern recognition.
668 (2013) 3414–3421
- 669 14. Wood, E., Baltrušaitis, T., Zhang, X., Sugano, Y., Robinson, P., Bulling, A.: Ren-
670 dering of eyes for eye-shape registration and gaze estimation. In: Proceedings of
671 the IEEE International Conference on Computer Vision. (2015) 3756–3764
- 672 15. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-synthesis for appearance-based
673 3d gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision
674 and Pattern Recognition. (2014) 1821–1828
- 675 16. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid
676 scene analysis. IEEE Transactions on pattern analysis and machine intelligence
677 20(11) (1998) 1254–1259

- 675 17. Wang, L., Lu, H., Ruan, X., Yang, M.H.: Deep networks for saliency detection via
676 local estimation and global search. In: Computer Vision and Pattern Recognition
677 (CVPR), 2015 IEEE Conference on, IEEE (2015) 3183–3192 678
- 678 18. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: Conference
679 on Computer Vision and Pattern Recognition. (2015) 680
- 680 19. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep
681 learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern
682 Recognition. (2015) 1265–1274 683
- 683 20. Gorji, S., Clark, J.J.: Attentional push: A deep convolutional network for aug-
684 menting image salience with shared attention modeling in social scenes. In: Pro-
685 ceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
686 (2017) 2510–2519 687
- 686 21. Benfold, B., Reid, I.: Guiding visual surveillance by tracking human attention. In:
687 British Machine Vision Conference. (September 2009) 688
- 688 22. Soo Park, H., Shi, J.: Social saliency prediction. In: Proceedings of the IEEE
689 Conference on Computer Vision and Pattern Recognition. (2015) 4777–4785 690
- 690 23. Cristani, M., Bazzani, L., Pagetti, G., Fossati, A., Tosato, D., Del Bue, A.,
691 Menegaz, G., Murino, V.: Social interaction discovery by statistical analysis of
692 f-formations. In: Proc. BMVC. (2011) 693
- 693 24. Chen, C.Y., Grauman, K.: Subjects and their objects: Localizing interactees for
694 a person-centric view of importance. International Journal of Computer Vision
695 (2016) 1–22 696
- 696 25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition.
arXiv preprint arXiv:1512.03385 (2015) 697
- 697 26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P.,
Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference
698 on computer vision, Springer (2014) 740–755 699
- 699 27. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The
700 pascal visual object classes (voc) challenge. International journal of computer
701 vision **88**(2) (2010) 303–338 702
- 702 28. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale
703 hierarchical image database. In: Computer Vision and Pattern Recognition, 2009.
704 CVPR 2009. IEEE Conference on, IEEE (2009) 248–255 705
- 705 29. : Pytorch: Tensors and dynamic neural networks in python with strong gpu ac-
celeration. <https://github.com/pytorch/pytorch> Accessed: 2017-11-03. 706
- 706 30. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans
707 look. In: Computer Vision, 2009 IEEE 12th international conference on, IEEE
708 (2009) 2106–2113 709
- 709 31. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It’s written all over your face: Full-
710 face appearance-based gaze estimation. In: Proc. IEEE International Conference
711 on Computer Vision and Pattern Recognition Workshops (CVPRW). (2017) 712
- 712 32. Hutman, T., Chela, M.K., Gillespie-Lynch, K., Sigman, M.: Selective visual atten-
713 tion at twelve months: Signs of autism in early social interactions. Journal of
714 autism and developmental disorders **42**(4) (2012) 487–498 715
- 715 33. Baltrušaitis, T., Robinson, P., Morency, L.P.: Openface: an open source facial
716 behavior analysis toolkit. In: Applications of Computer Vision (WACV), 2016
717 IEEE Winter Conference on, IEEE (2016) 1–10 718
- 718
- 719