

A Different Generative Approach to Ref-COD — Less Effective than GRCOD

A. Notes on Other Generative Models

We first review the Latent Diffusion Model used in our paper. It consists of an auto-encoder (VAE) and a UNet. The auto-encoder facilitates a two-way transformation between the RGB image $\mathbf{I}_c \in \mathbb{R}^{H \times W \times 3}$ and the latent space $\mathbf{z} \in \mathbb{R}^{h \times w \times c}$. Both the forward and backward processes of diffusion are carried out in the latent space, and we denote the noisy latent code at time t as $\mathbf{z}^{(t)} = \sqrt{\bar{\alpha}_t} \mathbf{z} + \sqrt{1 - \bar{\alpha}_t} \epsilon$, where $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$ is the noise schedule. β_s is the variance sampled from a variance schedule $\beta_t \in (0, 1)^T$. The UNet can be considered as a series of equally weighted denoiser $\epsilon_\theta(\mathbf{z}^{(t)}, t)$. The training objective \mathcal{L} can be simplified as:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}, \epsilon \sim \mathcal{N}(0, 1), t \in \mathcal{U}(T)} \left[\left\| \epsilon - \epsilon_\theta(\mathbf{z}^{(t)}, t) \right\|_2^2 \right] \quad (1)$$

Furthermore, to simplify comprehension and narration, we can reparametrize the output of UNet ϵ_θ as the form of v-predicton v_θ . The training objective can be further elaborated as:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}, \epsilon \sim \mathcal{N}(0, 1), t \in \mathcal{U}(T)} \left[\left\| \mathbf{z} - v_\theta(\mathbf{z}^{(t)}, t) \right\|_2^2 \right] \quad (2)$$

This implies that the goal of every training round is to denoise $\mathbf{z}^{(t)}$ to \mathbf{z} for any time step t .

Secondly, we present our task definition, using one-reference camouflaged object segmentation as an illustration. Given a data triplet $(\mathbf{I}_r, \mathbf{G}_r, \mathbf{I}_c)$, here \mathbf{I}_r and \mathbf{I}_c denote the support reference image and query camouflaged image respectively, both sharing an overlapping category c . \mathbf{G}_r is the mask of category c in the support reference image. Our task is to predict the mask corresponding to camouflaged category c in \mathbf{I}_c . In the strict one-reference camouflaged object segmentation setting, the category sets of the training set and the test set are disjoint.

Our objective is to fully utilize the priors in the Latent Diffusion Model and equip it with Few-reference Camouflaged Object Semantic Segmentation capabilities. This leads us to reuse the original VAE to convert $\mathbf{I}_r, \mathbf{I}_c$ and \mathbf{G}_r into latent variables $\mathbf{z}_r, \mathbf{z}_c$ and \mathbf{z}_p . Thus, our task is further simplified to explore how to improve the structure of UNet to v_θ^* so that it can accept $\mathbf{z}_r, \mathbf{z}_c$ and \mathbf{G}_r as inputs, and use \mathbf{z}_p as supervision.

This supervised approach in the latent space has been certified effective in tasks such as depth estimation and semantic segmentation. Concretely, our training objective \mathcal{L} is transformed into:

$$\mathcal{L} = \mathbb{E}_{(\mathbf{z}_r, \mathbf{z}_c, \mathbf{G}_r, \mathbf{z}_p) \sim \mathcal{D}} \left[\left\| \mathbf{z}_p - v_\theta^*(\mathbf{z}_r, \mathbf{z}_c, \mathbf{G}_r) \right\|_2^2 \right] \quad (3)$$

where \mathcal{D} represents the constructed training dataset. In addition, we omitted the input of time t . Our early experiments revealed that performing multiple steps of noise addition and denoising during training did not bring performance improvement.

B. Discussion

Limitations Our method, as the first diffusion-based Ref-COD model, proposes a simple and intuitive design, which maximizes the retention of the generative framework of LDM. There is still a lot of room for improvement in performance (especially in the n-reference Ref-COD setting), including more sophisticated model design and more optimized training strategies. We hope that our method can serve as a diffusion-based Ref-COD baseline to inspire more researchers to invest in this field.

On the other hand, we believe that our method is not limited to Ref-COD. Our framework has the potential to unify few-reference segmentation and open vocabulary segmentation by leveraging prompts from different modalities.

B.1. More details on generation process

In the above section we have discussed the generation process. In addition to the final choice of one-step image-to-mask sampling, we also tried multi-step noise-to-mask sampling and multi-step image-to-mask sampling. Here we detail the training objectives of these three generation processes.

one-step image-to-mask sampling We directly input the image and let the UNet output the mask. This process can be described as:

$$\mathcal{L}_1 = \mathbb{E}_{(\mathbf{z}_r, \mathbf{z}_c, \tilde{\mathbf{z}}_t, \mathbf{z}_p) \sim \mathcal{D}} \left[\left\| \mathbf{z}_p - v_\theta^*(\mathbf{z}_r, \mathbf{z}_c, \mathbf{z}_p) \right\|_2^2 \right] \quad (4)$$

Multi-step noise-to-mask generation We add noise to camouflaged mask \mathbf{z}_p , $\mathbf{z}_p^{(t)} = \sqrt{\bar{\alpha}_t} \mathbf{z}_p + \sqrt{1 - \bar{\alpha}_t} \epsilon$, and during inference we use $\mathbf{z}_p^{(0)}$ as the mask prediction. The supervised form is as follows:

$$\mathcal{L}_2 = \mathbb{E}_{(\mathbf{z}_r, \mathbf{z}_c, \mathbf{z}_p, \mathbf{z}_p) \sim \mathcal{D}, \epsilon \sim \mathcal{N}(0, 1), t \in \mathcal{U}(T)} \left[\left\| \mathbf{z}_p - v_\theta^*(\mathbf{z}_p^{(t)}, \mathbf{z}_r, \mathbf{z}_c, \mathbf{z}_p, t) \right\|_2^2 \right] \quad (5)$$

Multi-step image-to-mask generation We add image(as noise) to the camouflaged mask \mathbf{z}_p , $\mathbf{z}_p^{(t)} = \sqrt{\bar{\alpha}_t} \mathbf{z}_p + \sqrt{1 - \bar{\alpha}_t} \mathbf{z}_c$. The supervised form is as follows:

$$\mathcal{L}_3 = \mathbb{E}_{(\mathbf{z}_r, \mathbf{z}_c, \mathbf{z}_p, \mathbf{z}_p) \sim \mathcal{D}, t \in \mathcal{U}(T)} \left[\left\| \mathbf{z}_p - v_\theta^*(\mathbf{z}_p^{(t)}, \mathbf{z}_r, \mathbf{z}_c, \mathbf{z}_p, t) \right\|_2^2 \right] \quad (6)$$

B.2. Cross-attention tokenized interaction

In the paper, we only discussed how to inject information from the reference mask based on the reference fusion

108 self-attention mechanism method. Here we discuss how to
 109 inject information from the reference mask based on the To-
 110 kenized Interaction Cross-Attention method. There are also
 111 the following four ways.
 112

- 113 1. **Concatenation** We can convert the reference mask \mathbf{G}_r
 114 into an RGB image, encode I_r and \mathbf{G}_r into token se-
 115 quences using CLIP image encoder respectively, con-
 116 catenate them on the sequence, and finally use them as
 117 the input of cross-attention.
 118
- 119 2. **Multiplication** We can directly multiply \mathbf{G}_r on the
 120 image \mathbf{I}_r to form the image $\mathbf{I}_r^* = \mathbf{I}_r \cdot \mathbf{G}_r$, and finally
 121 encode \mathbf{I}_r^* into a token sequence using CLIP image en-
 122 coder as the input of cross-attention.
 123
- 124 3. **Addition** We can also directly add \mathbf{G}_r to the image
 125 \mathbf{I}_r to form the image $\mathbf{I}_r^* = 0.5\mathbf{I}_r + 0.5\mathbf{G}_r$. Similarly,
 126 we encode \mathbf{I}_r^* into a token sequence using CLIP image
 127 encoder as the input of cross-attention.
 128
- 129 4. **Attention Mask** We can use \mathbf{G}_r as an attention mask
 130 to control self-attention.
 131

132 To reduce the randomness caused by initial noise, enhance
 133 the influence of reference prompts, and ensure consistency
 134 between outputs and inputs, we employ classifier-free guid-
 135 ance (CFG) [?]. The query camouflaged latent z_c and con-
 136 dition τ are randomly set to null embedding with probability
 137 $p = 0.05$ in the training stage.
 138

We also adopt CFG in the inference stage. Specifically,
 139 our model outputs the $\tilde{z}_t(z_c, \tau)$ on the basis of three condi-
 140 tional outputs $\tilde{z}_t(z_c, \tau), \tilde{z}_t(\emptyset, \emptyset), \tilde{z}_t(z_c, \emptyset)$ (Eq. 7).
 141

$$\begin{aligned} \tilde{z}_t(z_c, \tau) &= \tilde{z}_t(\emptyset, \emptyset) \\ &+ \gamma_c \cdot (\tilde{z}_t(z_c, \emptyset) - \tilde{z}_t(\emptyset, \emptyset)) \\ &+ \gamma_\tau \cdot (\tilde{z}_t(z_c, \tau) - \tilde{z}_t(z_c, \emptyset)), \end{aligned} \quad (7)$$

142 where γ_q and γ_τ control the guidance of query camouflaged
 143 image and reference prompt, respectively.
 144

145 **Meta-architecture.** We investigate the three meta-
 146 architectures by applying different training strategies. In
 147 general, the model with all parameters trainable performs
 148 best. In addition, we study the effect of parameter-efficient
 149 fine-tuning on the models using LoRA. Performance degra-
 150 dation is observed for both architectures.
 151

152 When we further reduce the rank from 4 to 1, the per-
 153 formance of the one-step image-to-mask sampling slightly
 154 degrades. This phenomenon suggests that because SD was
 155 originally designed for generative tasks, its limited expres-
 156 sive capacity hinders transfer to segmentation tasks, and
 157 Multi-step generation is more sensitive to this characteris-
 158 tic.
 159

B.3. 1-reference to N-reference

160 So far, we have primarily explored the training and in-
 161 ference processes specifically designed for 1-reference sce-
 162 narios. A natural question arises: can this framework be
 163 extended to n-reference settings? To address this, we first
 164 present the simplest and most straightforward method for
 165 adaptation, which requires only minor modifications during
 166 the inference phase to accommodate n-reference tasks.
 167

168 In the paper, we introduced how to inject the informa-
 169 tion of the inference image into the features of the query
 170 camouflaged image using the reference fusion self-atten-
 171 tion mechanism. In inference, our support set R may contain
 172 more than one image, $S = \{I_{r1}, I_{r2}, \dots, I_{rn}\}$. We en-
 173 code each image into the features \mathbf{I}_{ri} . Correspondingly, af-
 174 ter mapping, we can obtain a series of $\mathbf{Q}_{ri}, \mathbf{K}_{ri}, \mathbf{V}_{ri}$ and
 175 $\mathbf{Q}_{ci}, \mathbf{K}_{ci}, \mathbf{V}_{ci}$. We can concatenate \mathbf{K}_{ci} and \mathbf{K}_{ci} to form
 176 $\mathbf{K}_{cr} = [\mathbf{K}_{ci}, \mathbf{K}_{r1}, \mathbf{K}_{r2}, \dots, \mathbf{K}_{rn}]$, and similarly we can ob-
 177 177
 178 178
 179 179
 180 180
 181 181
 182 182
 183 183
 184 184
 185 185
 186 186
 187 187
 188 188
 189 189
 190 190
 191 191
 192 192
 193 193
 194 194
 195 195
 196 196
 197 197
 198 198
 199 199
 200 200
 201 201
 202 202
 203 203
 204 204
 205 205
 206 206
 207 207
 208 208
 209 209
 210 210
 211 211
 212 212
 213 213
 214 214
 215 215

$$z_c^* = FusionA(z_c, z_r) = Att(Q_c, K_{cr}, V_{cr}) \quad (8)$$

192 While the aforementioned solutions enable N-reference
 193 inference, their performance does not match that of state-of-
 194 the-art (SOTA) models. This discrepancy primarily arises
 195 because the model receives only a single support image dur-
 196 ing the training phase, which leads to inconsistencies when
 197 transitioning to the inference phase with 5-reference or 10-
 198 reference configurations.
 199

200 To address this issue, we explore improvements from
 201 both the inference and training perspectives. From the per-
 202 spective of inference, transitioning from 1-reference to N-
 203 reference involves concatenating the keys and values of ad-
 204 dditional support samples, which significantly increases the
 205 number of keys and values processed during inference. To
 206 address this, we implement random sampling of the keys
 207 and values from the reference samples during inference, en-
 208 suring that their quantity matches that of the training phase.
 209 Another more straightforward idea is to introduce multiple
 210 reference samples during the training phase. In this way,
 211 the model can learn how to utilize multiple reference im-
 212 ages during training. we randomly select 1 to N reference
 213 samples as input using reference fusion self-attention dur-
 214 ing a single training iteration .Our experiments demonstrate
 215 that improvements during the training phase are more effec-
 216 tive than those during the inference phase.
 217

218 For cross-attention tokenized interaction, information
 219 of the reference mask can also be injected in the same
 220 four ways. There are just some slight differences in the
 221 implementation details. We carry out a comparison of
 222 two interaction methods paired with four injection meth-
 223 ods; these eight combinations are then verified experi-
 224 mentally. Overall, we observe that reference fusion self-
 225

attention (RFSA) mechanism outperforms Tokenized Interaction Cross-Attention (TCA). We attribute this mainly to the preservation and flexible utilization of information from the reference image by Ref-COD. Conversely, TCA, which only compresses support image to tokens via the CLIP image encoder, leads to some information loss. Notably, within the RFSA, the Concatenation method surpassed the other three. It offered a more free-form handling of RGB images and MASK information via subsequent learnable convolutional layers, compared to other hard injection methods. In the case of TCA, the Attention Mask method seems more apt as other operations are actually constrained by the CLIP image encoder. The CLIP image encoder itself is not good at dealing with mask information. Of course, we believe that there is still room for further exploration here, referring to FGVP.

B.4. Post processing

The original prediction of the model is an RGB three-channel image. We first average over the channel dimension to obtain a single-channel $\hat{\mathbf{G}}_c \in [0, 1]^{H \times W}$. Then we tried two thresholding methods, absolute threshold τ_a and relative threshold τ_r . The absolute threshold is a fixed value, and the final binary mask \mathbf{G}_c can be represented as:

$$\mathbf{G}_c = \begin{cases} 1, & \text{if } \hat{\mathbf{G}}_c > \tau_a \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

Using relative threshold, we have:

$$\mathbf{G}_c = \begin{cases} 1, & \text{if } \hat{\mathbf{G}}_c > \tau_r \max(\hat{\mathbf{G}}_c) \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

Table 1. Comparison of different thresholding methods

τ_r	0.2	0.25	0.3	0.35	0.4
mIoU	77.56	77.60	77.48	77.4	77.11
τ_a	0.1	0.15	0.2	0.25	0.3
mIoU	76.65	77.24	76.98	76.58	76

B.5. More ablation studies

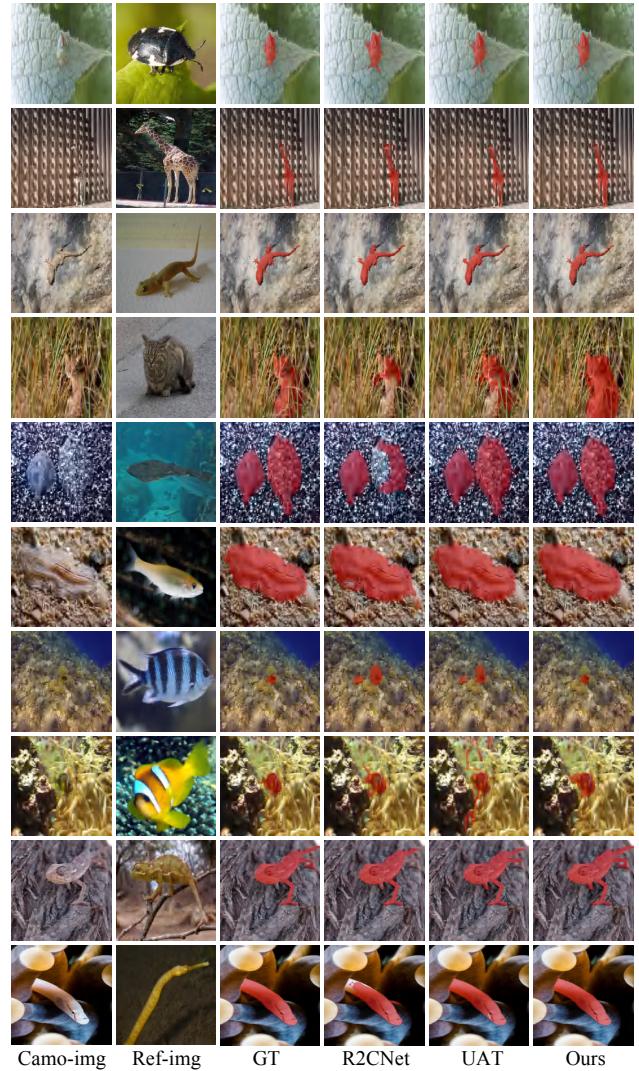
Multiplication We found in the experiment that Multiplication can be directly applied to RGB images, and another choice is to apply it to the latent space.

Table 2. Comparison of different Multiplication methods

Multiplication	mIoU
latent	62.11
RGB	63.18

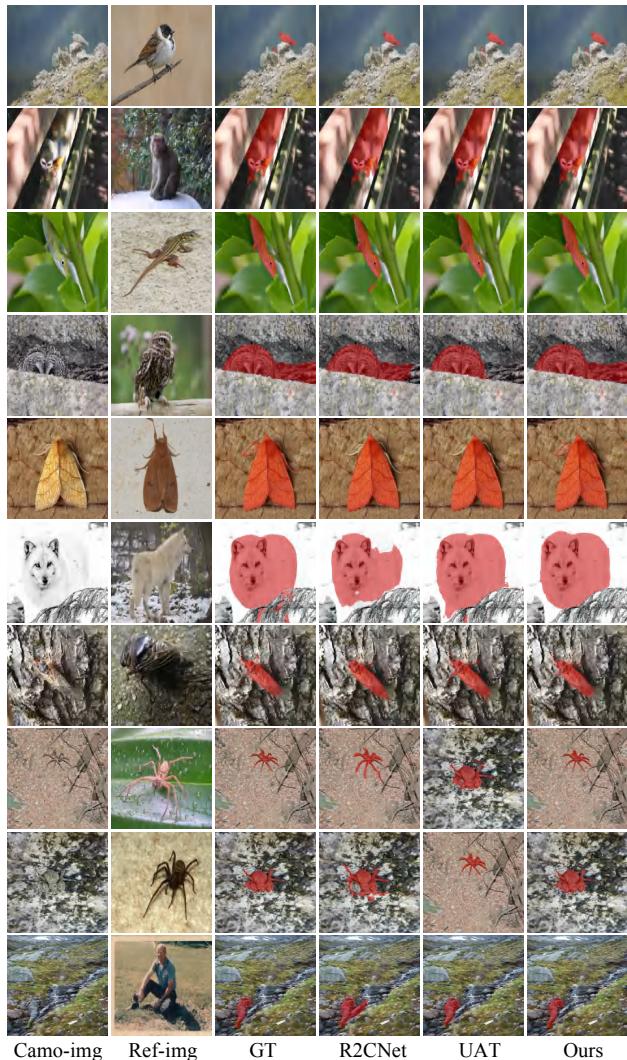
As shown in Tab. 2, the Multiplication method directly applied to RGB images achieved better results. However, the overall disparity is not significant.

B.6. Other visualization



216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
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

ICME 2025 Submission #575. Supplementary Material.

324
325
326
327
328
329
330
331
332
333

367

368

369

370

371

372

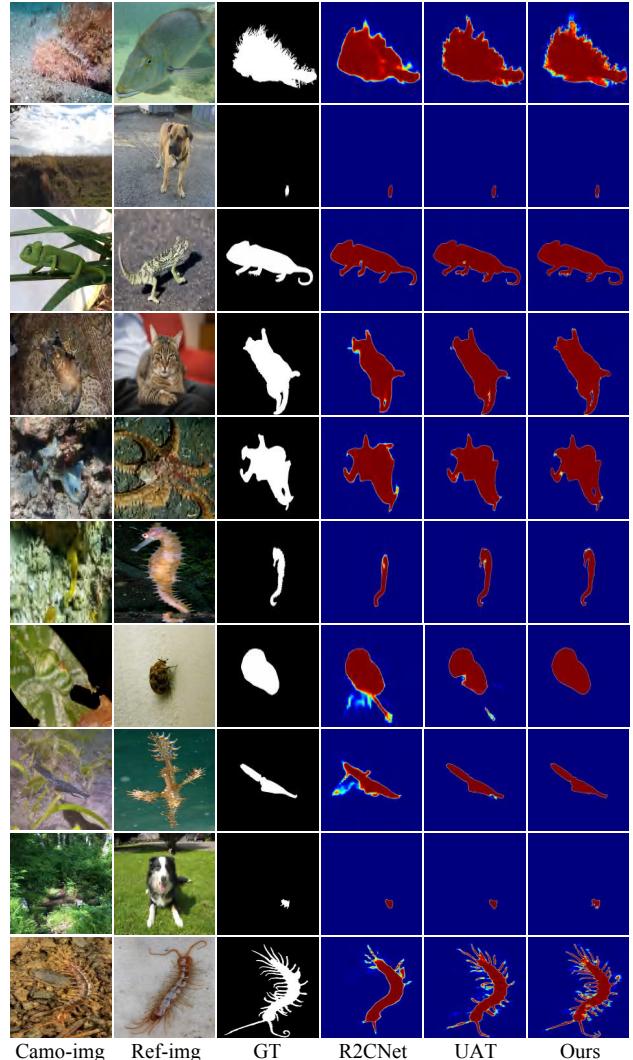
373

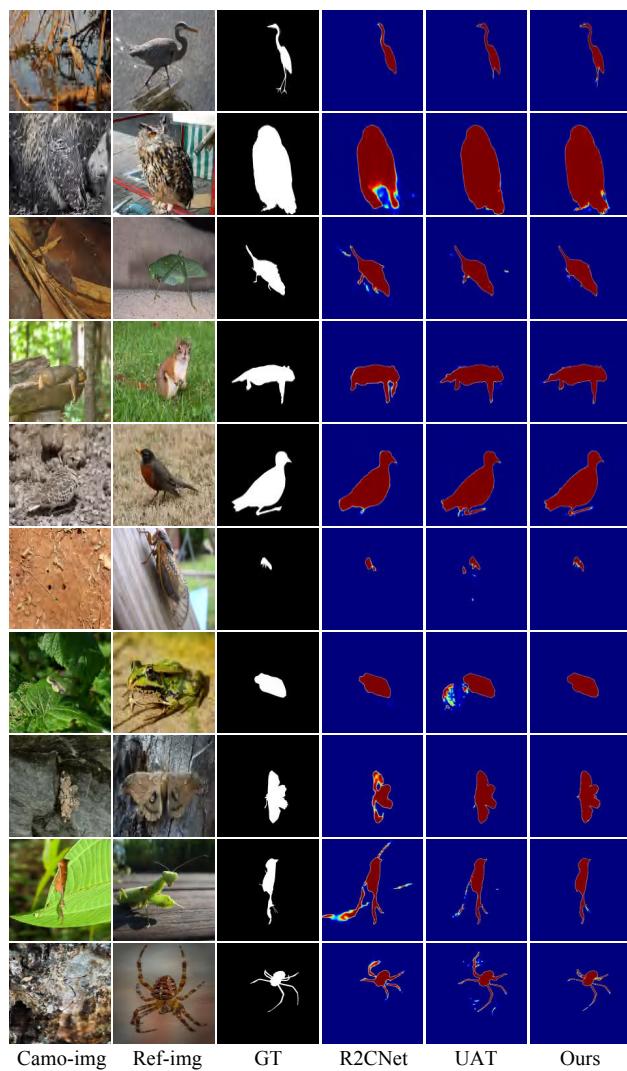
374

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



Camo-img Ref-img GT R2CNet UAT Ours

432 486
433 487
434 488
435 489
436 490
437 491
438 492
439 493
440 494
441 495
442 496
443 497
444 498
445 499
446 500
447 501
448 502
449 503
450 504
451 505
452 506
453 507
454 508
455 509
456 510
457 511
458 512
459 513
460 514
461 515
462 516
463 517
464 518
465 519
466 520
467 521
468 522
469 523
470 524
471 525
472 526
473 527
474 528
475 529
476 530
477 531
478 532
479 533
480 534
481 535
482 536
483 537
484 538
485 539