

A Supplementary material

Supplementary material is presented here as follows, first Section A.1 provides more extensive results showing that the histogram evaluation method successfully captures and shows the various statistical modalities related to the various pairings in the dataset. Section A.2, provides a further analysis on the conditionality of cGAN and the proposed method via carefully selected boundary cases which show the failure of the classic discriminator to learn conditionality. Section A.3 provides an additional evaluation of mode collapse for the depth prediction model. Section A.4 looks into the choice of weighting the different parts of the proposed loss function. Details are provided for reproducibility in Section A.5. Finally, an analysis of the training procedure is provided in Section A.6 to show that the training procedures did not encounter any degenerate situations.

A.1 Histogram evaluation criteria

The discriminator encodes the high dimensional space of input pairs into a lower dimensional latent space. Visualizing the empirical distribution on a high dimensional space is infeasible, however, since the encoded latent space learnt by the discriminator is compact, it is possible to visualize the empirical distribution by observing the latent space. This provides a direct insight into the discriminator performance and insight into the errors fed back for training the generator. For the purpose of this paper, a histogram is plotted in order to demonstrate the capacity of the discriminator to correctly classify underlying data distributions (see Figure A.1) as an example. A good cGAN discriminator should classify real-conditional as true, and all the three remaining pairs as fake even if the variables of the pairs are sampled from real data distribution (the case of a *contrario* real). Further analysis and results are provided to show the performance of the evaluation using the proposed histogram approach. As a reminder, 4 sets of data pairings are created as formalised in Section 2.1. The output response of the discriminator for each of these pairings is then plotted as a histogram with a different color. It can be noted that in the original GAN paper [14], the authors provide the intuition behind the underlying probability distributions. With the histogram test approach shown here, these distributions can be clearly observed.

The work presented in [14] showed that the optimal discriminator converges to:

$$D^* = \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y}|\mathbf{x}) + p_g(\mathbf{y}|\mathbf{x})} \quad (8)$$

Therefore considering the GAN cost function in Eq 1 with the definition in Eq 3 the cost function becomes:

$$V(G, D^*) = \min_G -\log(4) + 2D_{js}(p_g || p) \quad (9)$$

$p(\mathbf{x}, \mathbf{y}|\mathbf{x})$ is the real-conditional pair distribution while $p_g(\mathbf{x}, \mathbf{y}|\mathbf{x})$ is the generated-conditional pair. Therefore, Eq 1 corresponds to minimizing Jensen Shannon divergence between the probability distribution defined by the generator $p_g(\mathbf{x}, \mathbf{y}|\mathbf{x})$ and the real probability distribution $p(\mathbf{x}, \mathbf{y}|\mathbf{x})$. Therefore, for an optimal discriminator, the optimal generator is defined when $p_g(\mathbf{x}, \mathbf{y}|\mathbf{x})$ matches the real probability distribution $p(\mathbf{x}, \mathbf{y}|\mathbf{x})$. If the optimal discriminator fails to model conditionality, the generator may not be able to match the real probability distribution $p(\mathbf{x}, \mathbf{y}|\mathbf{x})$. That is why GAN models for semantic image synthesis suffer from poor image quality when trained with only adversarial supervision [48] and consider additional loss terms such as perceptual loss [23], L1[22] or feature-matching [45]. The segmentation-based discriminator proposed in [48] can be considered as a strong conditional discriminator by construction. The purpose of that paper was not explicitly enforcing conditionality, however, following the results presented in this paper, it could be concluded that they reach the

state of the art by their task specific error term which induces strong conditionality on the discriminator.

Figure A.1 shows the histograms of label-to-image and monocular depth prediction for epochs 20, 100 and 200. Several observations can be made:

- For the two tasks and for all evaluated epochs, the *a contrario* cGAN optimal discriminator correctly classifies each data pairing. The classical cGAN optimal discriminator fails by classifying real-*a-contrario* as true (greater than zero).
- For a *contrario* cGAN, during training, the conditional-generated samples are shifting towards towards the positive side of the function while training through the various epochs. This shows that the generator is correctly learning to minimize the divergence between the generated-conditional and real-conditional sets of data.
- Unlike classification metrics like accuracy, this histogram analysis provides more insight. For instance, even though the generated-conditional, generated-*a-contrario* and real-*a-contrario* are being classified as fake, three distributions are clearly distinguishable. There are therefore three different modalities for fake classification. There is only a single modality for the "true" classification which is clearly for real-conditional in yellow. The column showing the baseline optimal discriminator can be observed to model real-conditional and real-*a-contrario* with approximately the same distribution (i.e. only 1 mode is visible for both of these pairings). This indicates that the discriminator is invariant to the conditionality.

A.2 Conditionality analysis

In addition to Figure 1 in the paper, Figure A.2 is provided here to show various failure cases for the classic cGAN approach. The test was carried out on the Cityscapes dataset label-to-image trained as mentioned in Section 3. The histogram is plotted on the 500 Cityscapes where $\mathbf{y} \sim p(\mathbf{y}|\mathbf{x})$. The experiment was performed using the optimal discriminator for both models.

Extreme cases are chosen to assess the PatchGAN discriminator. Providing an "all-road", "all-car", "no-object" label for each pixel paired along the set of real images as an input pair. It can be seen clearly that PatchGAN discriminator classifies these pairs as "true". The *a contrario* cGAN successfully classifies these pairs as "fake". This suggests that the PatchGAN focuses only on \mathbf{y} instead of looking at the pair (\mathbf{x}, \mathbf{y}) .

Most PatchGAN-based methods do not pay careful attention to data pairing when training the discriminator, and subsequently the same conditional input image is reused for both real and generated input pairs in each mini-batch. More precisely, the conditional variable \mathbf{x} is always the same in each conditional pairing and only \mathbf{y} changes ($\mathbf{y} \sim p(\mathbf{y}|\mathbf{x})$ or $\mathbf{y} \sim p_g(\mathbf{y}|\mathbf{x})$). Subsequently the discriminator network learns to ignore the conditional input and the predicted true/fake label is only determined from the variable \mathbf{y} . During experimentation we tried to resolve this issue by not allowing the same \mathbf{x} to appear in both generated and real input pairings within a mini-batch. This test yielded the same results which suggest that ignoring the conditional variable is a fundamental problem of the classic PatchGAN architecture. As mentioned earlier, other architecture were also tested for conditionality and the result was the same. This suggest that that the problem is not specific to PatchGAN but generalises to cGAN architectures.

A.3 Mode collapse analysis

Mode collapse is the setting in which the generator learns to map several different inputs to the same output. A collapsing model is by construction unconditional. Only a few measures

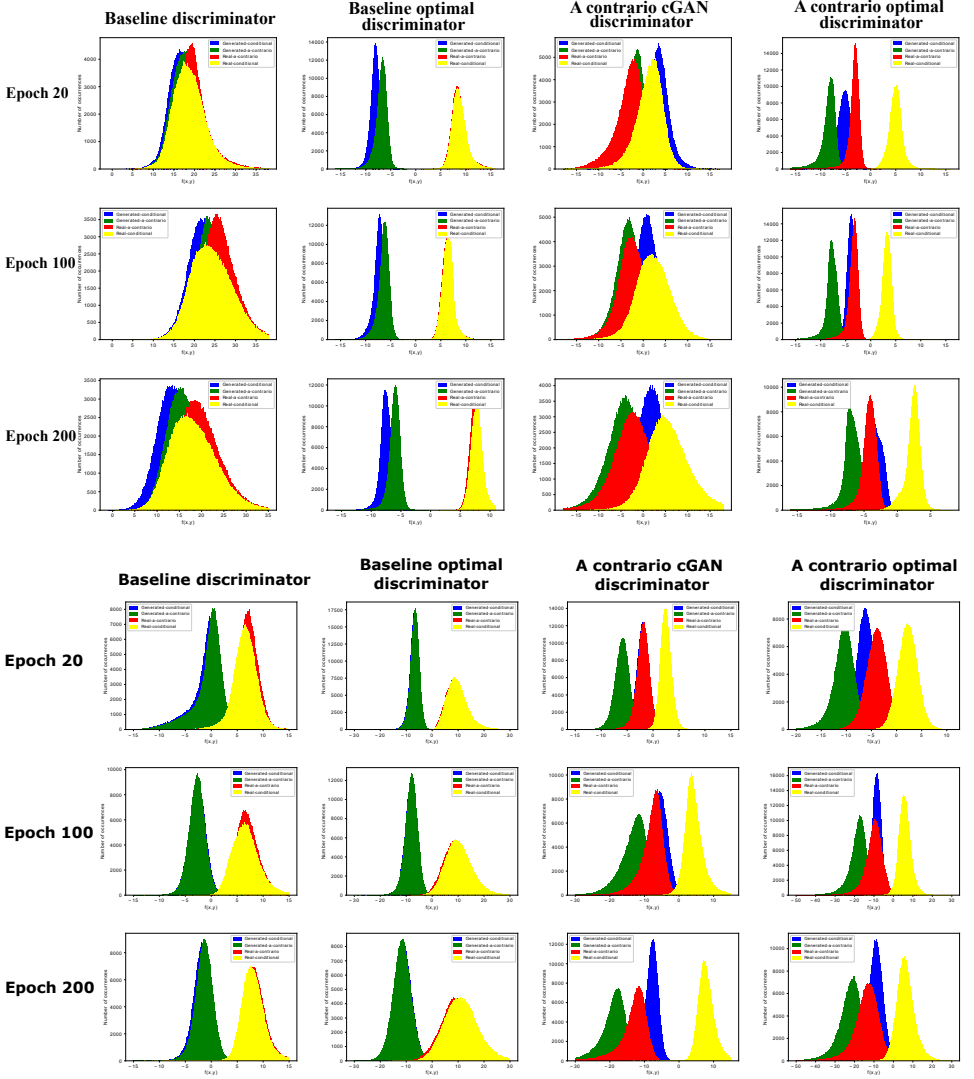


Figure A.1: Yellow and Blue are the classic data pairings being that of real-conditional and generated-conditional respectively. The proposed *a contrario* data pairings are Red and Green for real-*a-contrario* and generated-*a-contrario* respectively. The last convolution layer $f(x,y)$ shows positive values for "true" and negative values for "false" since the Sigmoid activation was used for training. The histogram evaluation is performed with dropout and batch normalization during training. The top three rows show Cityscape label-to-image histograms. The bottom three rows show the NYU Depth monocular depth prediction histograms. The *a contrario* cGAN optimal discriminator correctly classifies each data pairing while the PatchGAN optimal discriminator fails by classifying real-*a-contrario* as true (greater than zero).

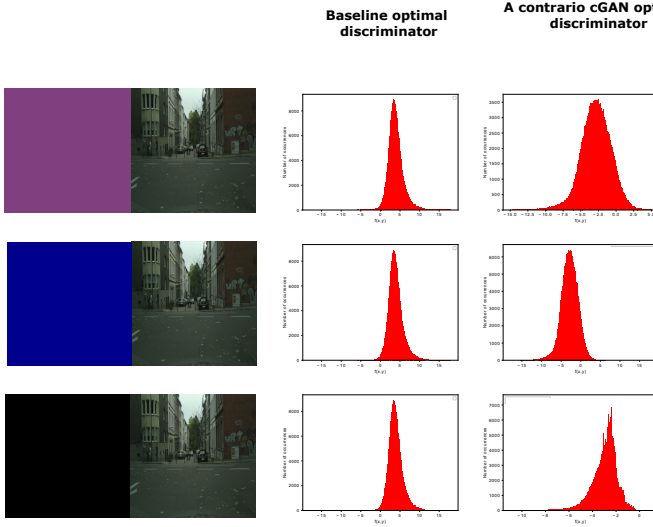


Figure A.2: Extreme cases are provided to evaluate the two discriminators. The first row represents the "all-road" label for all the 500 validation images. The second row represents the "all-car" label. The third row represents the "no-object" label. It can be seen clearly that the PatchGAN discriminator fails to classify these pairs as fake while *a contrario* cGAN succeeds to classify them correctly.

have been designed to explicitly evaluate this issue [3, 44, 55]. MS-SSIM [55, 56] measures a multi-scale structural similarity index and birthday paradox [3] concerns the probability that, in a set of n randomly chosen outputs, some pair of them will be duplicates. Another approach, NDB [44], presents a simple method to evaluate generative models based on relative proportions of samples that fall into predetermined bins.

The analysis provided in this section is an extension of the experiments done on depth prediction. Figure A.3 shows the evolution of the NDB measure over training iterations using the NDB score (the less, the better) for both pix2pix baseline and *a contrario* cGAN models trained on the NYU Depth V2 training set [37]. Out of the 12 trained models, the best model (in terms of RMSE log) is chosen for the evaluation. For clustering and evaluating NDB, non overlapping patches of 64×64 are considered. At the end of the training the NDB/ k ($k = 100$) of the *a contrario* cGAN is 0.550 while the baseline achieves only 0.645. This indicates that *a contrario* model **generalizes better**. This is also observed qualitatively in Figure A.8. Training with the counter examples helps the discriminator to model conditionality. Thus, the generator search space is restricted to only conditional space. The generator is penalized for non-conditionality even if the generation is realistic.

A.4 Loss function analysis

An ablation study on Eq 6 was performed. Each term that contributes to the adversarial loss is weighted by λ_i . Eq 6 becomes:

$$\begin{aligned} \mathcal{L}_{adv} = \min_G \max_D & \left[\lambda_1 \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}), \mathbf{y} \sim p(\mathbf{y}|\mathbf{x})} [\log(D(\mathbf{x}, \mathbf{y}))] + \lambda_2 \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} [\log(1 - D(\mathbf{x}, G(\mathbf{x})))] \right] + \\ & \max_D \left[\lambda_3 \mathbb{E}_{\tilde{\mathbf{x}} \sim p(\tilde{\mathbf{x}}), \mathbf{y} \sim p(\mathbf{y})} [\log(1 - D(\tilde{\mathbf{x}}, \mathbf{y}))] + \lambda_4 \mathbb{E}_{\tilde{\mathbf{x}} \sim p(\tilde{\mathbf{x}}), \mathbf{x} \sim p(\mathbf{x})} [\log(1 - D(\tilde{\mathbf{x}}, G(\mathbf{x})))] \right] \end{aligned} \quad (10)$$

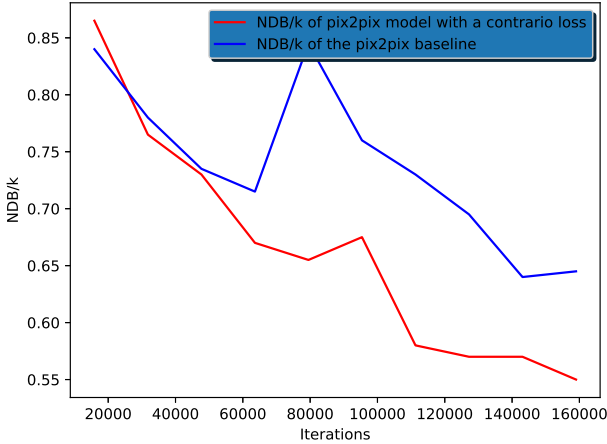


Figure A.3: An analysis of mode collapse using the NDB criteria (lower values are better) throughout training on the NYU depthV2 dataset. It can be concluded from this evaluation that the proposed approach is much better at avoiding mode collapse due to the restricted search space of the generator.

Three strategies were considered for the weighting. The models were trained on the Cityscapes label-to-image dataset with the same settings described earlier (Section 3.2). Figure A.4 shows the mIoU for different *a contrario* cGAN models trained with different choices for λ_i .

- **Strategy 1:** Equal contribution for each term : $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4$.
- **Strategy 2:** Balancing the "fake" and "true" contributions. Since there are 3 data pairings classified as fake and only 1 real pair as true, equal balancing of true/fake gives: $\lambda_1 = 1, \lambda_2 = \lambda_3 = \lambda_4 = 0.33$
- **Strategy 3:** Testing the significance of both *a contrario* error terms for fake and real images. In this case only 3 terms with real-a-contrario is tested : $\lambda_1 = \lambda_2 = \lambda_3 = 0.5, \lambda_4 = 0$.

In this simple test, Strategy 1 gives the best results. Strategy 2 seems less stable. Strategy 3 succeeds to learn conditionality, however, it may not capture conditionality for generated images during training. Each of these strategies succeed to model conditionality, however, Strategy 1 converges faster and yields a better final result in terms of mIOU.

A.5 Reproducibility

Various experiments were performed using different datasets and input-output modalities. Some extra detail is provided here for reproducibility purposes. In all the experiments using the pix2pix baseline, random jitter was applied by resizing the 256×256 input images to 286×286 and then randomly cropping back to size 256×256 . All networks were trained from scratch. Weights were initialized from a Gaussian distribution with mean 0 and standard deviation 0.02. The Adam optimizer was used with a learning rate of 0.0002, and momentum parameters $\beta_1 = 0.5, \beta_2 = 0.999$. A linear decay is applied starting from epoch 100, reaching 0 at epoch 200. Dropout is used during training. As in the original implementation [22], the discriminator is a PatchGan with a receptive field of 70×70 . Similarly pix2pixHD [54],

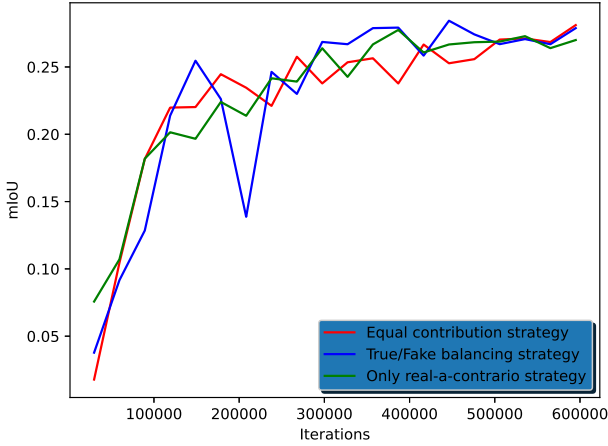


Figure A.4: The mIoU evaluation for different choice of λ_i . The strategy 1 of giving equal contribution yield the best results. However, there is no major difference on the convergence or the performances at epoch 200 between the different strategies

SPADE [41] and CC-FPSE [33] were trained with the same hyper-parameters as mentioned in their respective papers. For label-to-image, a U-Net256 with skip connections was used for the generator. A U-Net with 9 ResNet blocks was used for depth prediction, the last channel is 1 instead of 3 and the activation of the last convolution layer generator is *Relu* instead of *Tanh*.

For the image-to-label task, a U-Net256 with skip connections was used for the generator but the output channel size was chosen to be 19 instead of 3 for segmentation of 19 classes. The activation of the last convolution layer of the generator was changed to a softmax to predict class probability for segmentation purposes.

A.6 Training details

Figure A.5(a) shows the gradient of the classic and proposed *a contrario* cGANs trained on Cityscapes [7] label-to-image with and without *a contrario* (see Section 3.2). The mean absolute value of the gradient is reported in order to demonstrate the stability of the training. Neither vanishing nor exploding gradient is observed for both models. Figure A.5(b) shows the training loss of the optimal discriminator trained as described in Section 3.1 for both models with the generator fixed at epoch 200. Both models converge rapidly to 0. Allowing the discriminator to converge for one epoch is enough to obtain the optimal discriminator with a fixed generator.

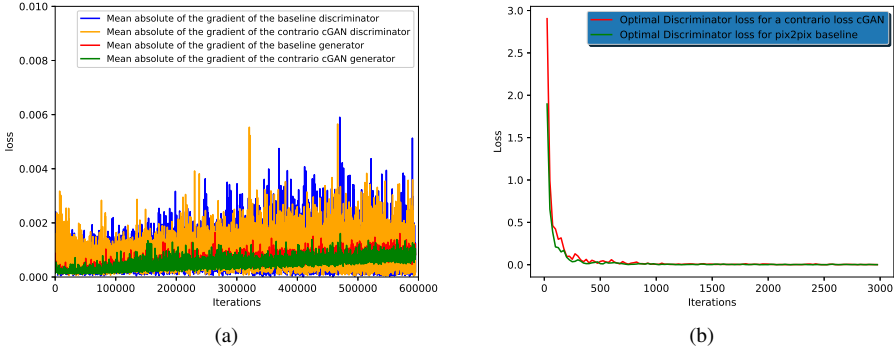


Figure A.5: (a) The mean absolute value of the gradients of the generator and discriminator for both baseline and a *contrario* cGAN models trained on Cityscapes[7]. The gradient is stable and it is neither vanishing nor exploding. (b) The loss function of the optimal discriminators when the generator is fixed. Both losses converge rapidly to 0.

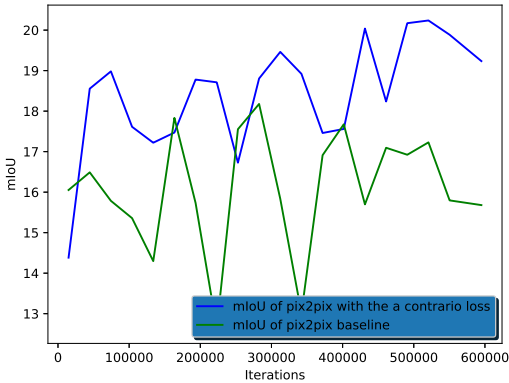


Figure A.6: mIoU for the Cityscape image-to-label dataset throughout training. The proposed method consistently obtains more accurate results and finishes with a largely different score at the end of training 19.23 versus for the baseline 15.97.

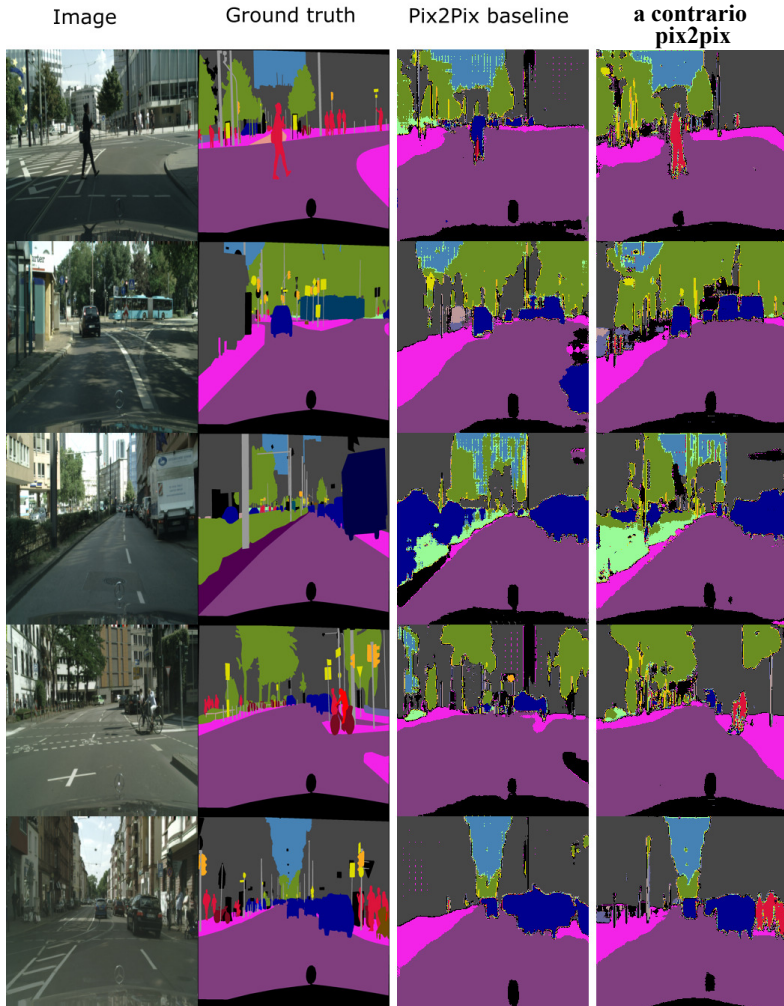


Figure A.7: Qualitative results of Cityscape image-to-label task. It can be seen that the baseline model hallucinates objects. For instance, in the second row, the baseline hallucinates cars while the *a contrario* cGAN segments the scene better. In the first row, the baseline wrongly classifies the pedestrian as a car. While training the model, the discriminator does not penalize the generator for these miss-classifications

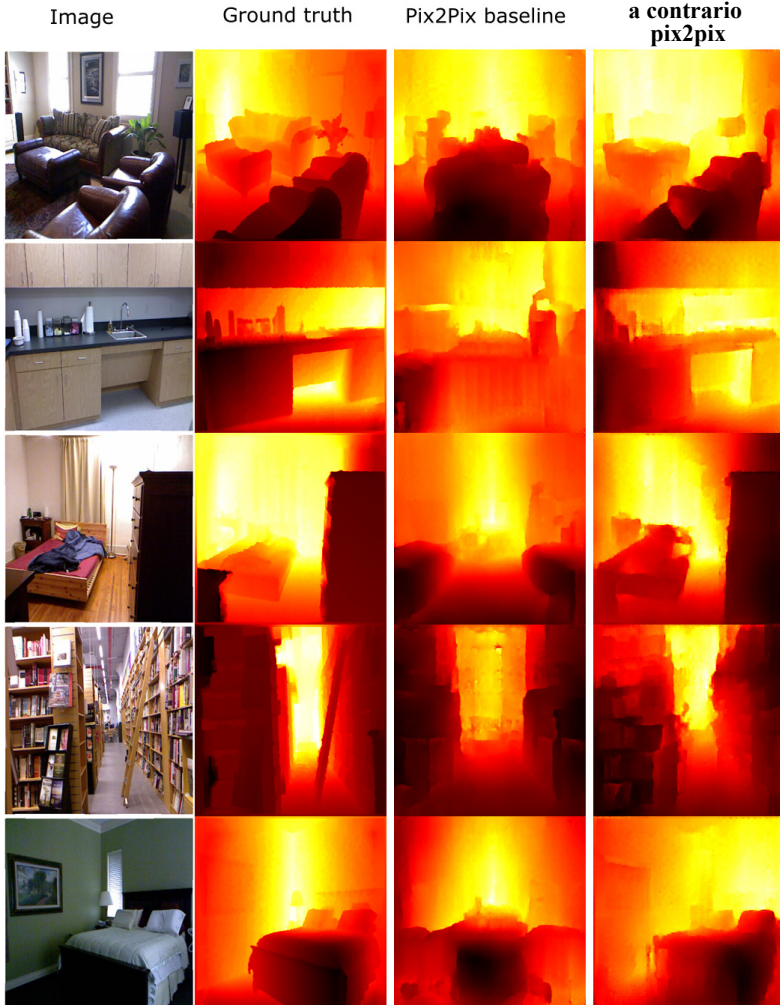


Figure A.8: Qualitative results for depth prediction. The *a contrario* cGAN shows better performance and more consistent prediction with respect to the input. The first row shows a case of mode collapse for the baseline as it ignores completely the input.

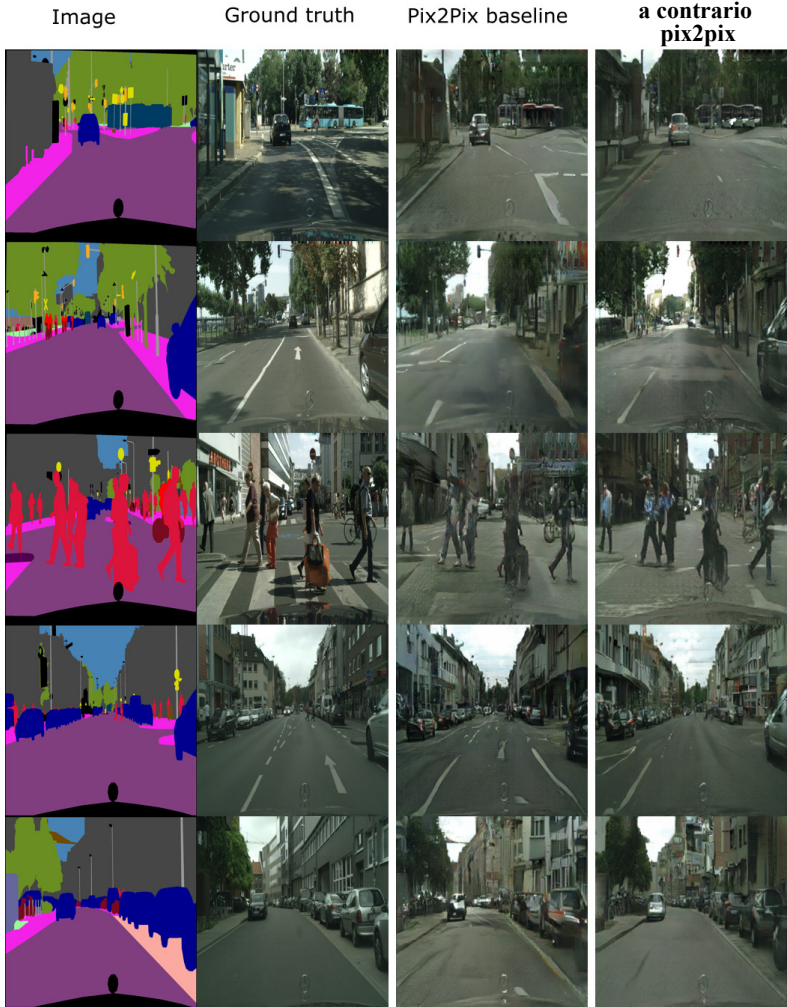


Figure A.9: Qualitative results of Cityscapes label-to-image synthesis. In line with the quantitative results reported in Section 3.2, the qualitative results show better results for the *a contrario* in comparison to the baseline.



Figure A.10: Qualitative comparison between different state-of-the-art methods for label-to-image trained and tested on Cityscapes[7] dataset. As observed, CC-FPSE baseline is the best baseline among classic cGAN. The *a contrario* improves all the baseline and the best model among the 6 models is *a contrario* CC-FPSE