Internal Diverse Image Completion
Supplementary Material

Noa Alkobi
Technion

Tamar Rott Shaham
Technion, MIT

Tomer Michaeli
Technion

1. Training

Models and training time All the generators and discriminators are composed of 5 convolutional blocks consisting of `conv-BN-LeakyReLU`. The last block of the discriminator does not include an activation layer and the activation layer of the last block in the generator is Tanh instead of LeakyReLU. The slope of the Leaky-ReLU for negative values is 0.2. Our models (both generator and discriminator) consist of 32 channels in the first 4 scales. Every 4 scales the number of channels is increased by a factor of 2. At scales where the number of channels is modified (first and every 4th scale) we initialize the convolution weights and biases with normal distribution with mean=0 and std=0.02. At all other scales, the weights and biases are initialized with those from the previous scale. Training takes about 2 hours on a Quadro RTX 8000 GPU for an image size of 193×256 and generating a new sample at inference takes less than a second per image.

Coarse scales At the coarse scales, the "real" image presented to the discriminator is a naively inpainted version of the masked input image, and therefore no masking is performed within the discriminator. Each of those coarse scales is trained for 2000 iterations using the Adam optimizer [1]. The learning rate for both the generator and the discriminator is $5 \cdot 10^{-4}$ (which decreases after 1600 iterations by 0.1) and the momentum parameters are $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Training is done with two losses: (i) WGAN-GP loss with gradient penalty weight of $\lambda = 0.1$ (ii) Reconstruction loss with weight of $\alpha = 10$. In each iteration we perform three gradient steps for updating D followed by three gradient steps for updating G.

Fine scales For scales at which there exist enough valid patches, the "real" image presented to the discriminator is the masked input image, and masking is performed within the discriminator so as to ignore all patches containing masked pixels. Each such fine scale is trained for 3000 iterations using the Adam optimizer [1] with a learning rate of $5 \cdot 10^{-5}$ for both the generator and discriminator, with the same momentum parameters as the coarser scales. We construct $\tilde{m}_n$, the soft version of the mask m_n at the n-th scale, by dilating m_n with a disc of size $\min(N - n, 5)$ pixels and convolving the result with a Gaussian filter with $\sigma = 5$. The soft mask is then forced to be zeros at the invalid pixels of the image. Training is done with the same losses as in the coarse scales. We adjust the losses weights as follows: (i) The gradient penalty weight is $\delta \cdot \lambda$ where δ is the ratio between the number of pixels within the image and the number of valid pixels and $\lambda = 0.1$. (ii) The reconstruction loss weight is calculated according to $\delta_{rec} \cdot \alpha$ where δ_{rec} is the ratio between the number of pixels in the image and the sum of all elements in $\tilde{m}_n$. Here we take $\alpha = 10$ for the coarser scales and we increase it to 100 in the last third of the scales.

Naive completion Obtaining the naive completion is inspired by DIP [3]. We train a U-Net network with depth depending on the image size, for 1000 iterations using the Adam optimizer with a learning rate of 10^{-3}. Besides the losses we use, which are described in the main text, all the rest is done as in [3]. Examples of naive inpainting results are presented in Fig. S1.

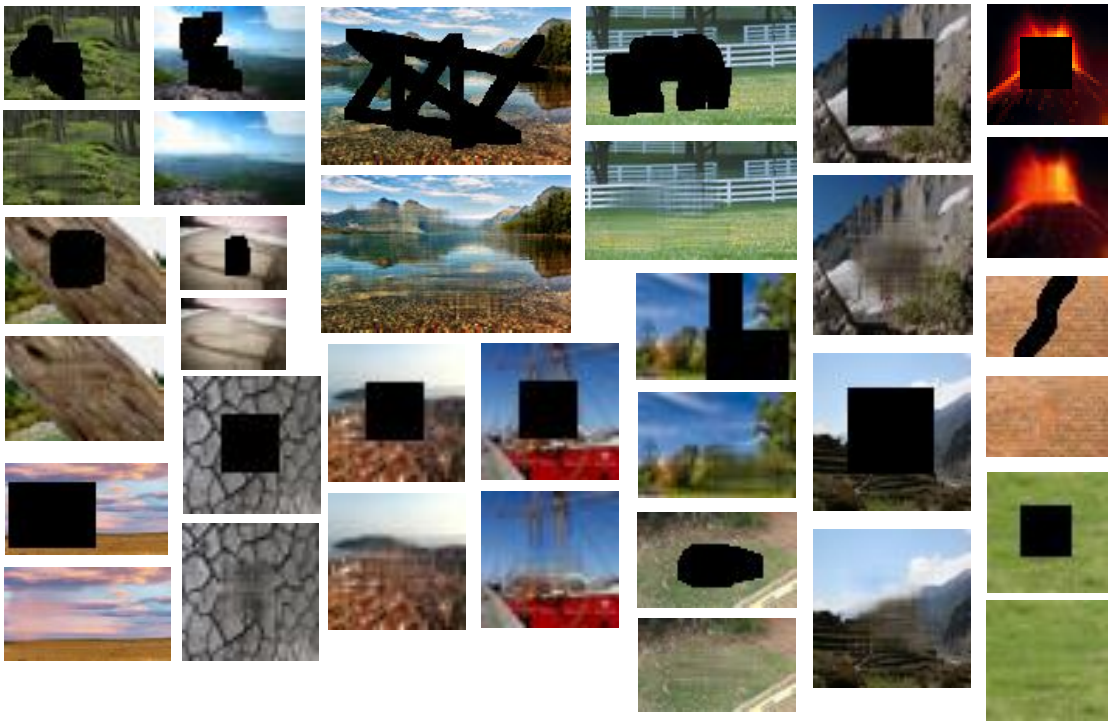


Figure S1. **Examples for the naive completions used as "real" images at coarse scales.** At coarse scales, we complete the (low resolution version of the) input image using the modified DIP framework. This provides a reasonable estimate for the global structure at the missing region.

2. Controlling the diversity

For editing applications, users may wish to control the level of diversity. In our method there are two ways of controlling the diversity of the results: (i) by enlarging the standard deviation of the noise sampled during training, and (ii) by enlarging the area of the new noise sampled at test time.

Enlarging the standard deviation of the injected noise We inject to the generator in all scales zero mean Gaussian noise with unit variance 1, which we multiply by a gain calculated according to the missing details in each scale. In order to check the influence of the standard deviation on the diversity of our method, we trained three different models on the same image. Each model was trained with a different noise variance (1,2, and 4). For each model, we calculated the semantic diversity (LPIPS) over 1000 pairs of samples. As can be seen in Tab. S1, enlarging the variance of the noise has a relatively small effect on the diversity of our method.

	LPIPS - Semantic Diversity
$z \sim \mathcal{N}(0,1)$	$102 \cdot 10^{-6}$
$z \sim \mathcal{N}(0,2)$	$129 \cdot 10^{-6}$
$z \sim \mathcal{N}(0,4)$	$128 \cdot 10^{-6}$

Table S1. **Different noise sampled.** LPIPS is calculated between 1000 pairs of samples on a specific image. Enlarging the variance of the noise barely affects the diversity of our method.

Enlarging the area of the new noise sampled An alternative way of controlling diversity is by changing the area of the new noise sampled at test time. Sampling a larger area in each pyramid scale results in higher diversity, but may also modify the valid area of the image. This trade-off is demonstrated in Fig. S2. Here, high diversity refers to the setting where the new noise fills the entire mask area, without erosion. Medium diversity refers to partially eroding the mask (by less than half a receptive-field). Note that in our normal diversity setting, we intentionally do not erode half of the full receptive field (which includes the up-sampling operation), in order to allow increased diversity. This results in only minor changes to the regions outside the mask (see STD image). The semantic diversity of our method with these three modes is shown in Tab. S2.

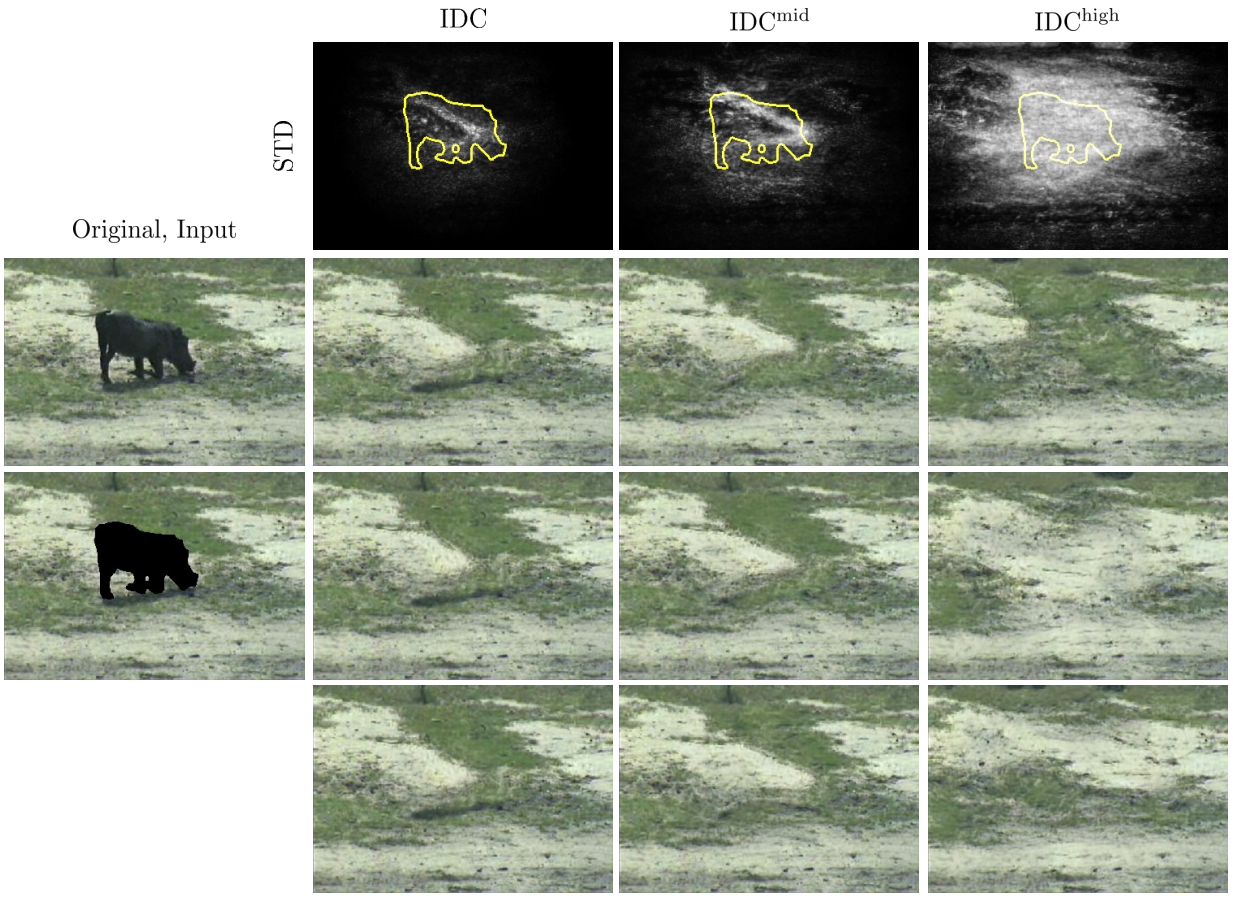


Figure S2. **Diversity control.** We can control the diversity of the completions by changing the area into which we inject new noise samples. In our normal setting, only pixels inside the mask are changed, while enlarging the diversity change a bigger area around the mask. The first row represents the STD over 20 samples for each degree of diversity. Note that the STD in the first column (low diversity) leaks a bit over the mask, however, it is negligible.

	LPIPS - Semantic Diversity
IDC	$15 \cdot 10^{-3}$
IDC$^{\text{mid}}$	$40 \cdot 10^{-3}$
IDC$^{\text{high}}$	$212 \cdot 10^{-3}$

Table S2. **Controlling the diversity.** We report semantic diversity of our completions on Part-Imagenet, with the three variants shown in Fig. S2. Letting our method change areas outside the mask results in higher diversity.

3. Additional results

Figure S3 shows inpainting results for the image presented in Fig. 1 in the main text, where here we use the DSI [2] and ICT [4] models that were trained on the Places dataset, rather than the ones trained on Imagenet (CoMod-GAN does not have an ImageNet model). One can see that with these models as well, DSI and ICT introduce artifacts. This is while our results are more realistic. Additional examples for our diverse completions are presented in Fig. S4. Figure S5 presents the results of other models, for the images presented in Fig. 6 in the main text.

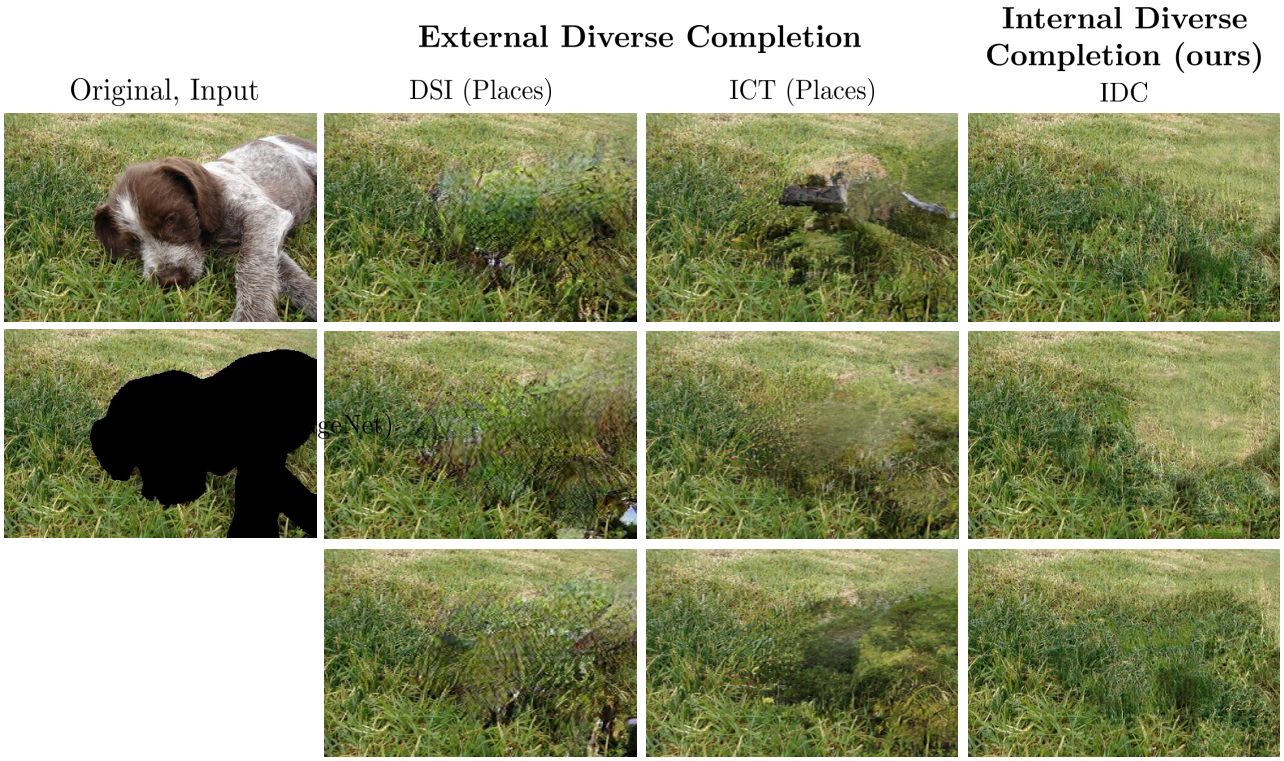


Figure S3. **Diverse Internal Completion.** Here we compare our method to DSI and ICT models trained on Places. Similar to its ImageNet counterparts (shown in Fig. 1 of the main text), those models as well generate artifacts. Our model, on the other hand, outputs photo-realistic results.

4. Importance of the Coarse Scales

The first row of Fig. S6 shows the results of our method when omitting the coarser scales in the model. Namely, we start the training from the first masked scale, $i-1$. For comparison, the last row shows our full training scheme. As can be seen, the coarse scales are crucial for preserving global structures within the image.

5. Importance of Masking the BN Layer

At the fine scales, where we do perform masking, the discriminator needs to ignore all invalid pixels in the real image. This is done by masking the elements of the discrimination map that correspond to patches containing invalid pixels. In a convolutional model without BN layers, this operation guarantees that the discriminator is not affected by the missing region. However, in the presence of BN layers, this is not enough because the BN statistics (mean and variance) are computed over their entire input feature map. Therefore, it is also crucial to compute the BN statistics in each layer only over valid features (not affected by invalid pixels). Figure S7 illustrates the importance of passing through the BN layer only valid patches. One can see that when not masking the BN layer, the results are blurry and contain artifacts.

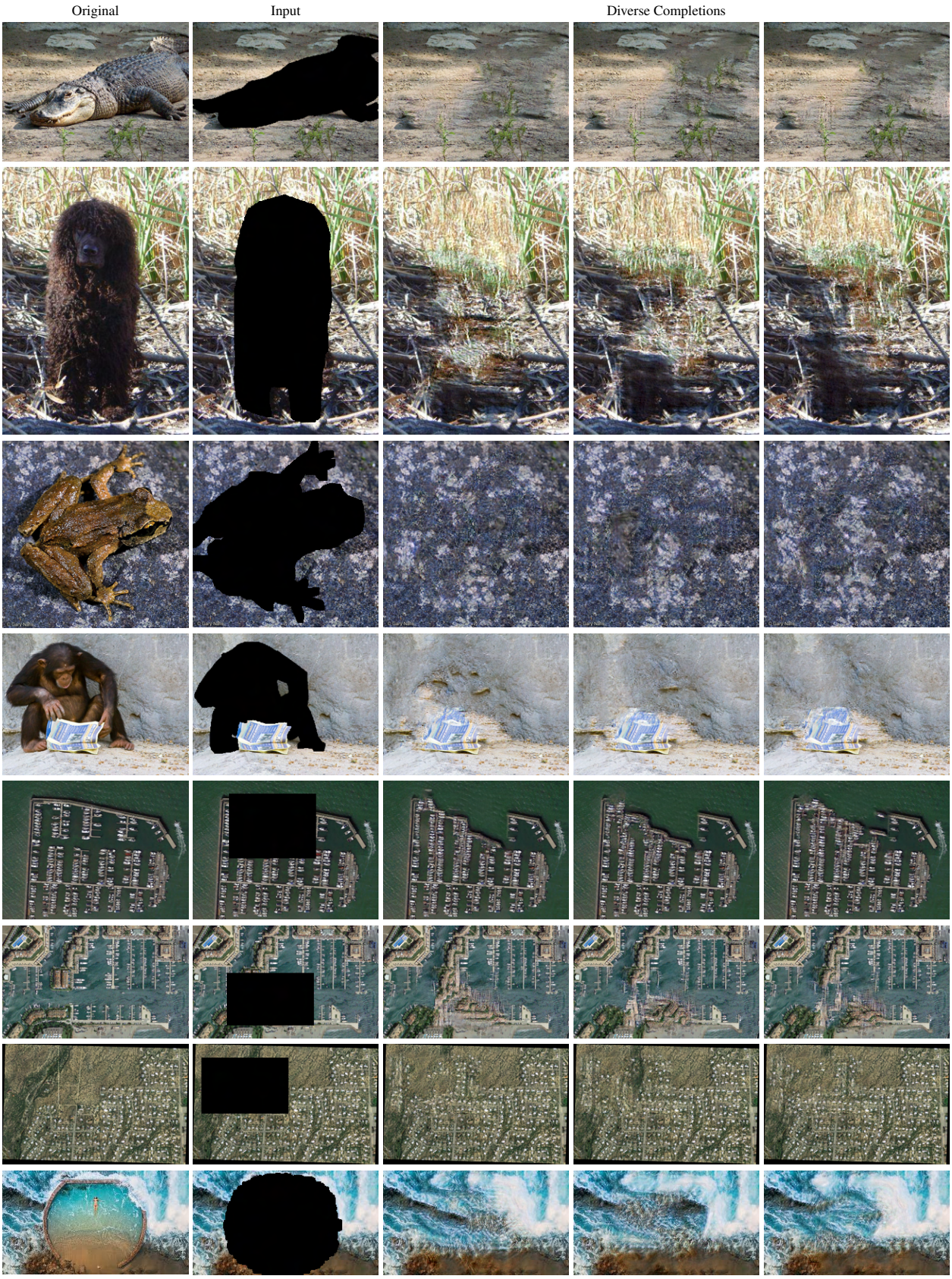


Figure S4. **Diverse completions of our method.**

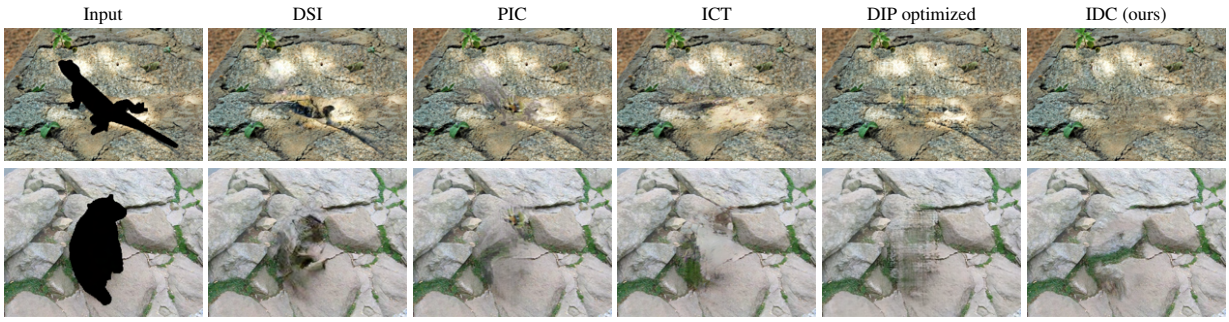


Figure S5. **Qualitative comparison.** Results of other models, for the images of Fig. 6 in the main text. For external methods, we used models trained on Places. Our method is at least comparable to baselines in terms of visual quality, while it is the only one that both offers diversity and is applicable to arbitrary domains (being internally trained).

Figure S6. **Importance of coarse scales.** When our model contains no coarse scales, the global structure of the image is damaged (sky in the trees). In contrast, the full model (having coarse scales learned with a naive completion and without masking) manages to preserve the global structure of the image.

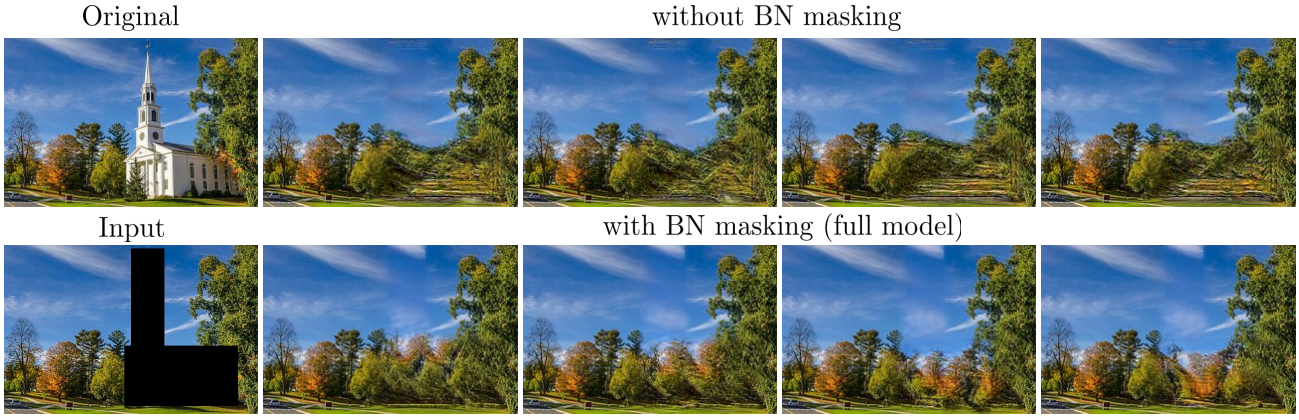


Figure S7. **Importance of BN masking.** The first row shows completions obtained when masking only the output of the discriminator. This results in blur and artifacts. The second row presents our method, which also applies BN masking. This leads to more realistic completions.

6. Limitation

In the setting of inpainting, the mask may hide parts of semantic objects, which internal methods cannot complete, as presented in Fig. S8. However, this task is challenging even for external methods.

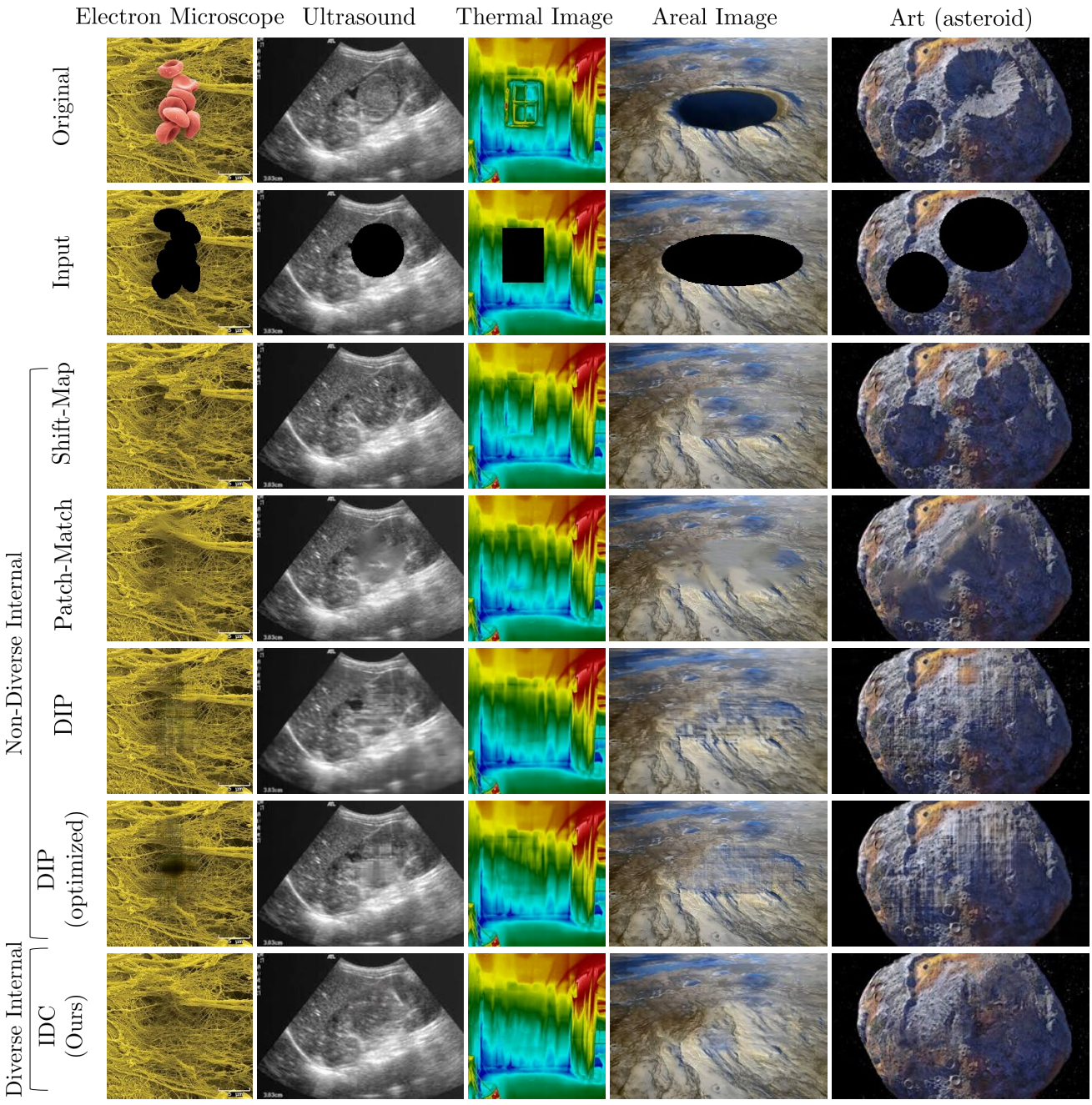
Figure S8. **Limitation.** When the missing region includes part of a semantic object (*e.g.* the woman's legs), our model fails to provide plausible completions. Note, however, that this task is challenging even for externally trained methods.

7. Patch-Match

A comment is in place regarding Patch-Match, which uses a stochastic algorithm for finding approximate nearest neighbor patches. While this algorithm should theoretically converge to the same solution despite its randomness, in practice different runs result in slightly different completions. Superficially, this could suggest that Patch-Match should also be regarded as an internal diverse method. However, in practice, the diversity arising from Patch-Match's randomness is actually quite low. For example, for 128×128 masks, Patch-Match's LPIPS diversity is 35×10^{-3} while ours is 60×10^{-3}. We therefore do not regard this unintentional byproduct of the algorithm's implementation as genuine diversity.

8. Completion of images from other domains

Figure S9 shows completion results for the images of Fig. 2 in the main text, obtained using competing methods. One can see that external methods often insert objects to the images.



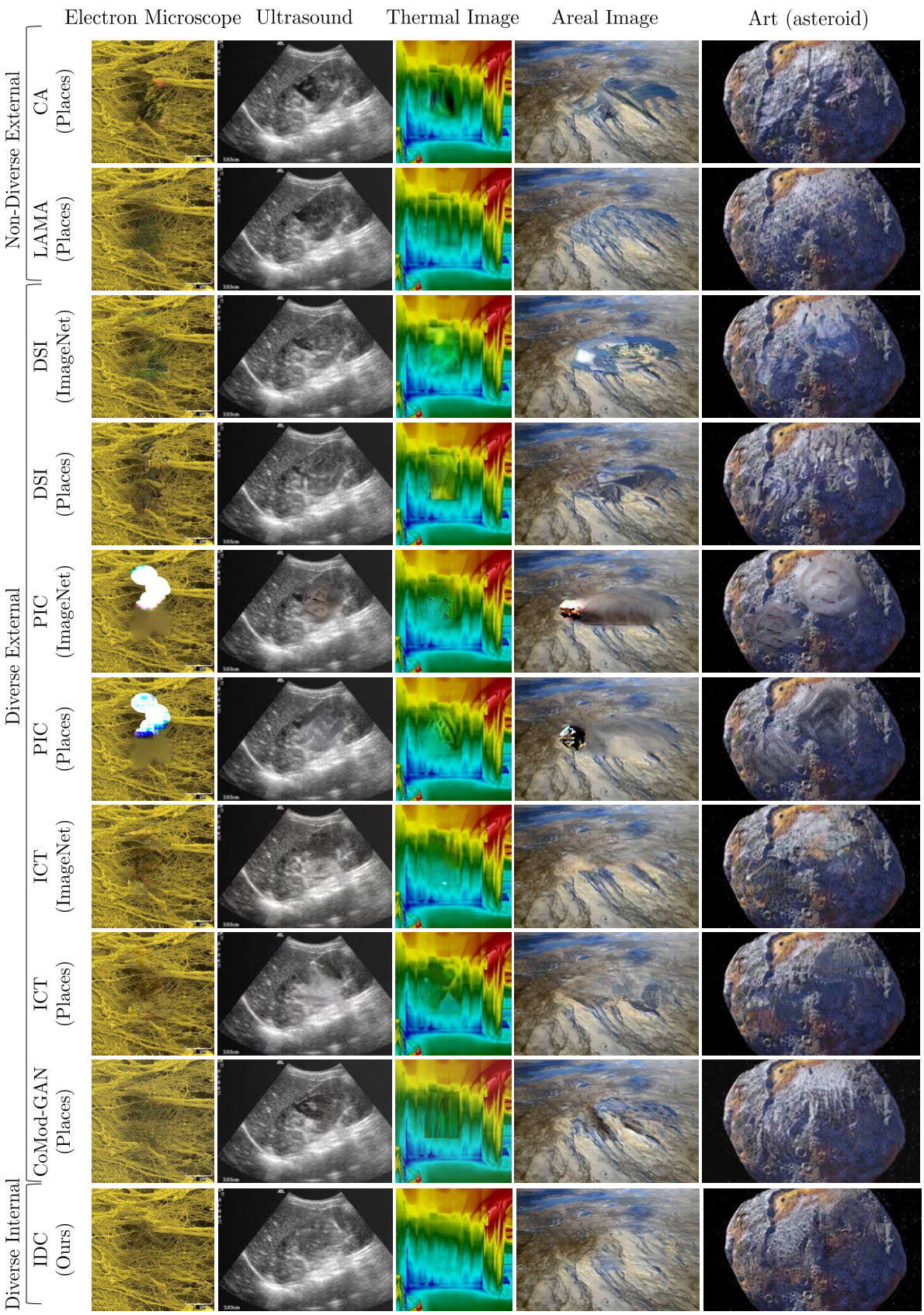


Figure S9. **Completion with all methods for other domains.**

9. Modified DIP

Recall that in the coarse scales of our model, we use naive inpainting, which we obtain from a modified variant of the DIP method. Figure S10 shows comparisons between the original DIP algorithm and our modified DIP. One can see the importance of adding the color consistency loss into to original DIP. Our modified DIP scheme was optimized for low-resolution images (we only used it for the coarse pyramid scales). Using it 'as is' on the full resolution images does not improve upon the original DIP.

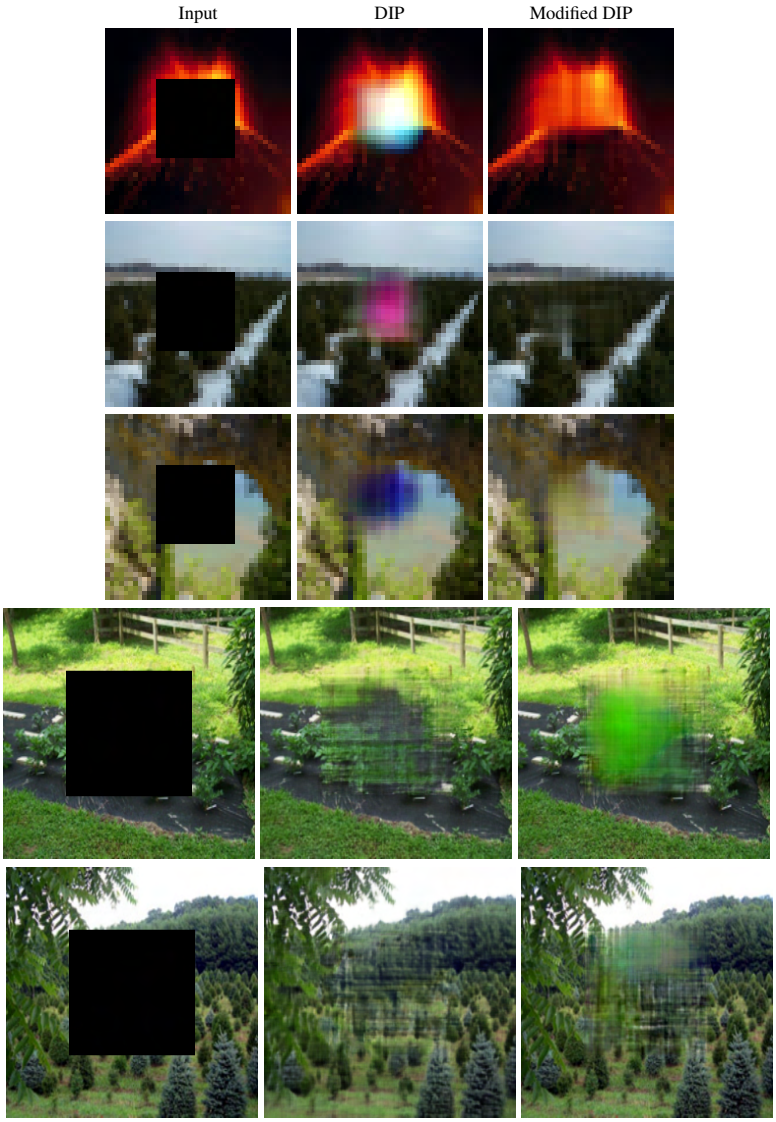


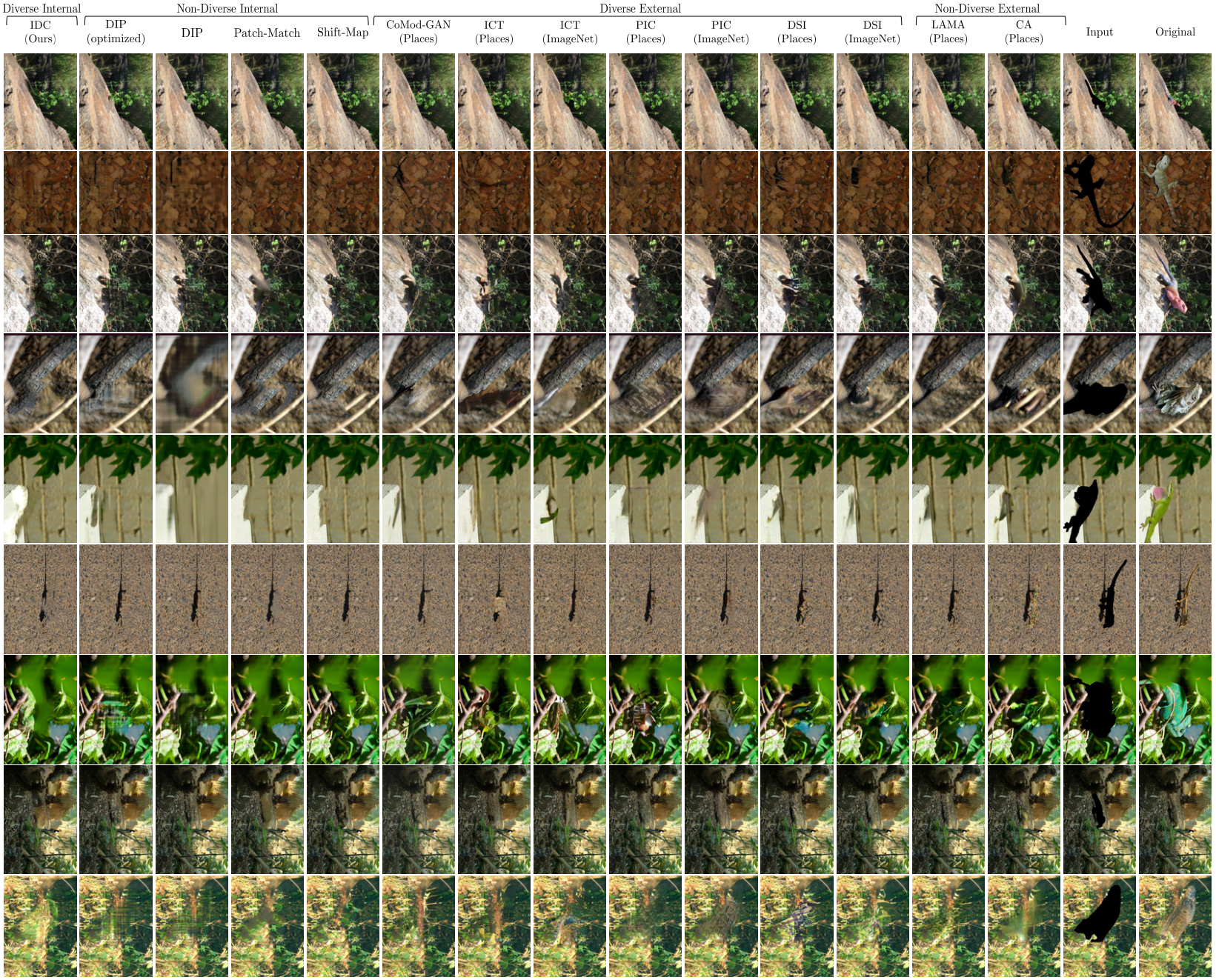
Figure S10. **Original DIP vs. our modified DIP.** Our modified DIP is better in low resolution images as can be seen in the first three rows. However, when using it on high resolution images (last two rows) it does not necessarily improve the results.

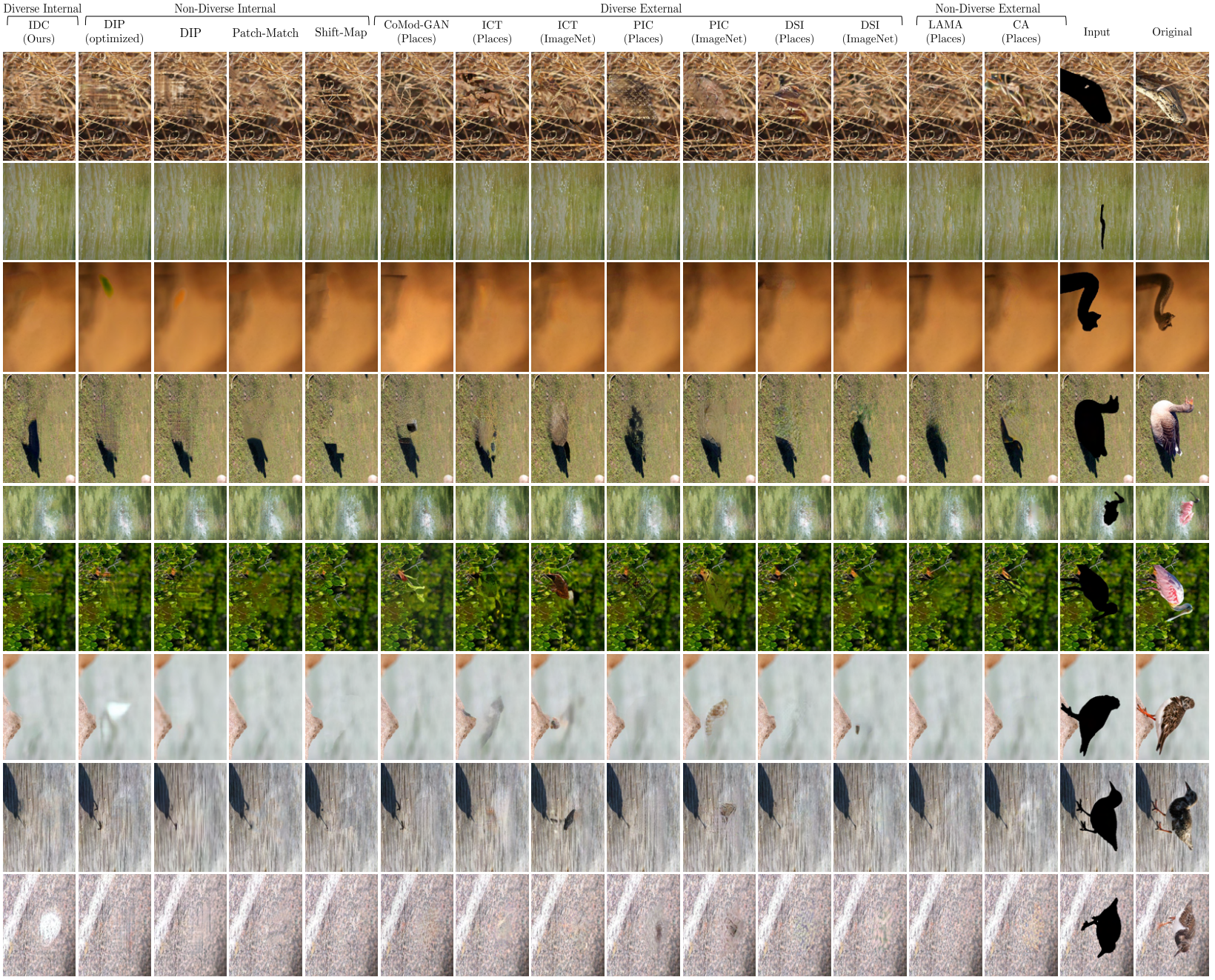
10. AMT survey

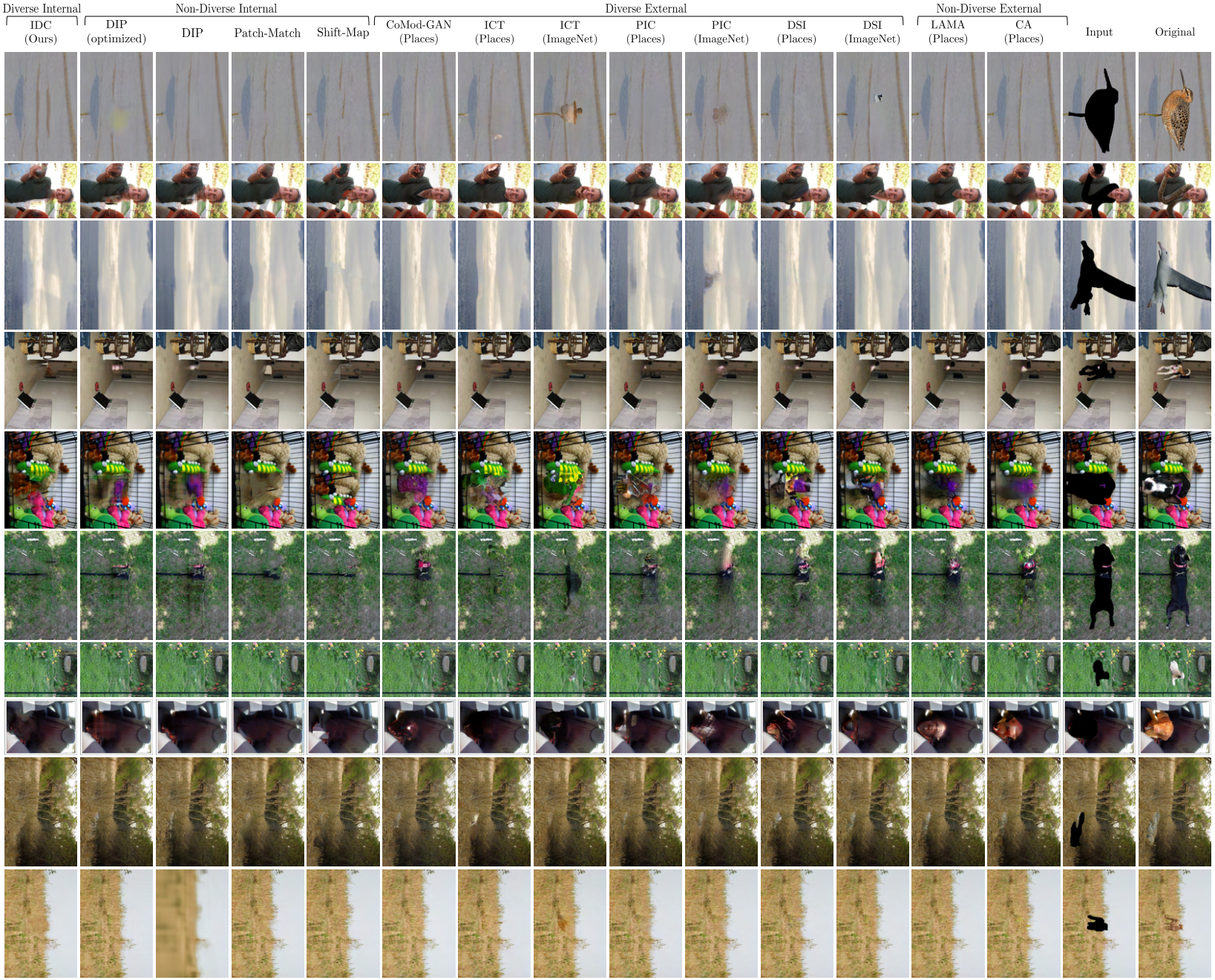
We start the survey with 5 tutorial questions with feedback. In the tutorial questions, it is clear which completion is better. Figure S11 depicts the images that participated in the tutorial. After the tutorial, the user was presented with 45 questions. In each of them, we display the original image, the masked image, our completion and one competing completion. All images presented in the user study are shown in Fig S12.

Figure S11. **AMT tutorial images.** We present to the user a short tutorial before answering the test. Users selected the completion they preferred and received feedback.


Diverse Internal
Non-Diverse Internal
Diverse External
Non-Diverse External
IDC (Ours)
DIP (optimized)
DIP
Patch-Match
Shift-Map
CoMod-GAN (Places)
ICT (Places)
ICT (ImageNet)
PIC (Places)
PIC (ImageNet)
DSI (Places)
DSI (ImageNet)
LAMA (Places)
CA (Places)
Input
Original


Diverse Internal
Non-Diverse Internal
Diverse External
Non-Diverse External
IDC (Ours)
DIP (optimized)
DIP
Patch-Match
Shift-Map
CoMod-GAN (Places)
ICT (Places)
ICT (ImageNet)
PIC (Places)
PIC (ImageNet)
DSI (Places)
DSI (ImageNet)
LAMA (Places)
CA (Places)
Input
Original


Diverse Internal
Non-Diverse Internal
Diverse External
Non-Diverse External
IDC (Ours)
DIP (optimized)
DIP
Patch-Match
Shift-Map
CoMod-GAN (Places)
ICT (Places)
ICT (ImageNet)
PIC (Places)
PIC (ImageNet)
DSI (Places)
DSI (ImageNet)
LAMA (Places)
CA (Places)
Input
Original


Diverse Internal
Non-Diverse Internal
Diverse External
Non-Diverse External
IDC (Ours)
DIP (optimized)
DIP
Patch-Match
Shift-Map
CoMod-GAN (Places)
ICT (Places)
ICT (ImageNet)
PIC (Places)
PIC (ImageNet)
DSI (Places)
DSI (ImageNet)
LAMA (Places)
CA (Places)
Input
Original

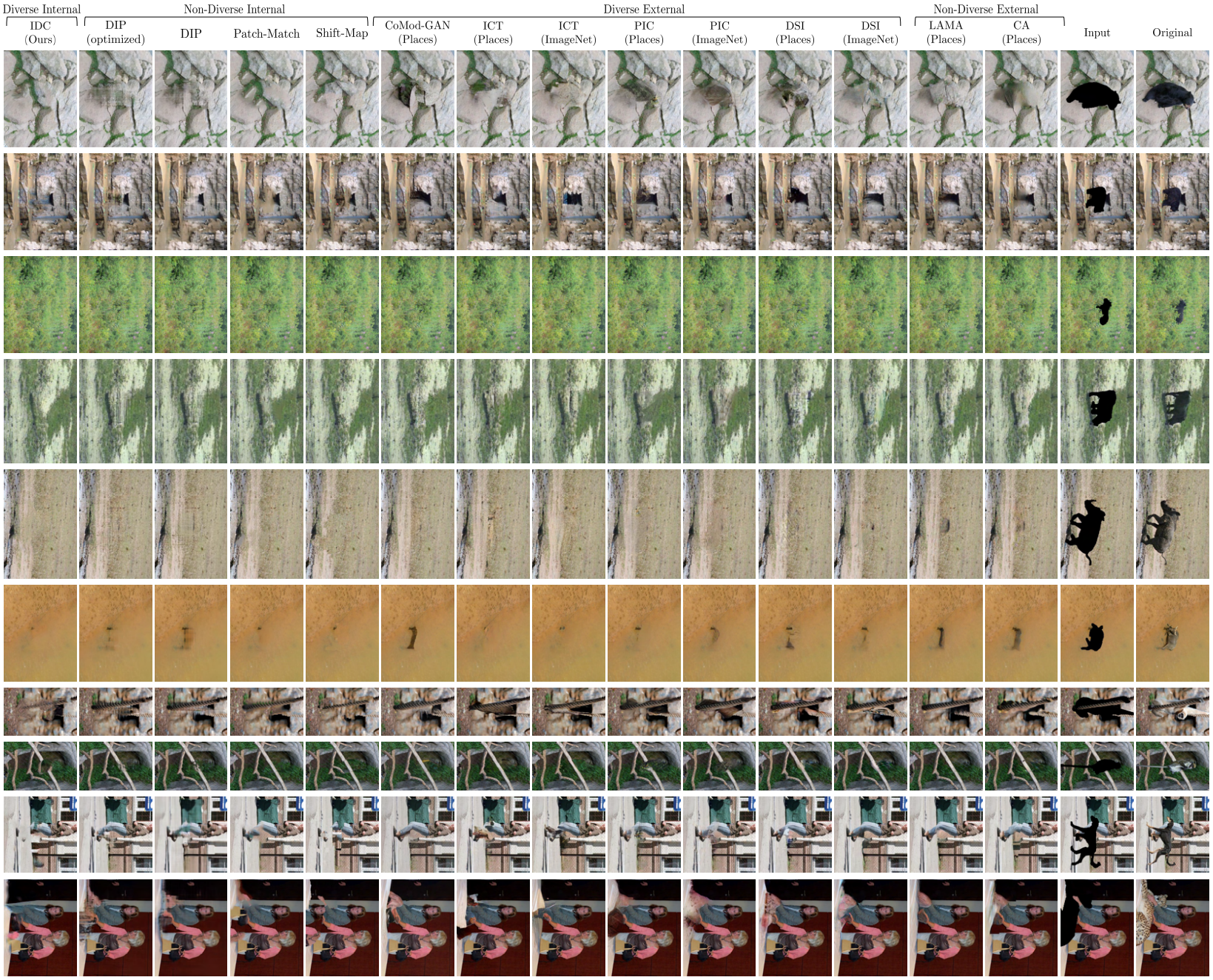

Diverse Internal
Non-Diverse Internal
Diverse External
Non-Diverse External
IDC (Ours)
DIP (optimized)
DIP
Patch-Match
Shift-Map
CoMod-GAN (Places)
ICT (Places)
ICT (ImageNet)
PIC (Places)
PIC (ImageNet)
DSI (Places)
DSI (ImageNet)
LAMA (Places)
CA (Places)
Input
Original

Figure S12. **Images used in our user study.**

11. Places validation images

For comparison to baselines on the task of inpainting with arbitrary masks (not necessarily hiding a whole semantic object), we used 50 images from the Places dataset, presented in Fig. S13.

Figure S13. **Places validation dataset.** We randomly chose 50 images from the Places validation set, from the subcategories Mountains, Hills, Desert and Sky. Images were resized to 256×256 since some of the baselines accept only images having this size.

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[2] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10775–10784, 2021.

[3] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018.

[4] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4692–4701, 2021.