Robust Learning Through Cross-Task Consistency
Supplementary Material

http://consistency.epfl.ch

Abstract

The supplementary material provides qualitative and quantitative results and method details which were moved out of the main paper in interest of space. Specifically:

I. A video evaluation showing various networks applied frame-by-frame to a Youtube video (§ 1).

II. A live demo where you can apply our networks to your own query images (§ 2).

III. Results of consistency with unsupervised tasks (§ 3).

IV. Discussion of how the perceptual loss formulation handles ill-posed tasks and associated experiments (§ 4).

V. Description of our strategy for balancing different loss terms (§ 5).

VI. Plots showing that minimizing the direct term does not minimize consistency (§ 6).

VII. Derivation of generic path length criterion (§ 7).

VIII. Sensitivity analysis: Comparison of different edge selection strategies (§ 8).

IX. Sensitivity analysis: Training with consistency over multiple path lengths (§ 9).

X. Results with standard error over multiple initialization seeds (§ 10).

XI. Results on *NYUv2* dataset (§ 11).

XII. Results on *Taskonomy* dataset reported with *additional perceptual tasks* (§ 12.2) and using common *task-specific metrics* (e.g., mean and median angular error for surface normals) (§ 12.1).

XIII. More qualitative results provided in § 13.1 and on the project website.

XIV. More qualitative results on estimating surface normal out of middle domains before and after enforcing consistency. (§ 13.2).

XV. Definition and visualizations of the "Blind Guess" (statistically informed guess) for the Taskonomy dataset (§ 14).

XVI. Code: including pretrained models, runnable examples, and a Docker (§ 15).

1. Video Evaluation

The project website includes video clips that provide various results from the proposed framework as well as the baselines. In particular, the video clips provide results of different stages of the method applied on the YouTube video used in [7], which we find insightful. We recommend watching the clips.

2. Live Demo

The project website includes a live demo that allows you to upload your own query images. The demo will run that query through our servers and return the results, so that you can visualize the predictions for various tasks from networks trained with and without consistency, as well as compare to baselines. The demo page also contains a link to the "demo archive" where you can browse uploads from other users.

3. Consistency with Unsupervised Tasks

As described in the main paper, the tasks a network is constrained to be consistent with could be *unsupervised* or *self-supervised* too. This is particularly useful if the dataset in hand is either single-task (as most common datasets are) or includes few tasks. Unsupervised tasks allow generating new domains without any additional supervision, thus enable utilizing denser cross-task consistency constraints during training. Examples of such tasks are 2D texture edges and 2D keypoints (SURF[1]), which were included in our dictionary.[1] Such tasks have fixed operators thus can be applied on any images to produce their respective domains with no additional supervision.

Interestingly, enforcing cross-task consistency with such unsupervised domains, even without using any supervised ones, still led to better fitting the data and producing improved predictions. Fig. 1 shows the results of learning to predict surface normals, while the two consistency domains were 2D texture edges and 2D keypoints. In other words, using the notation $x \rightarrow y_1 \rightarrow Y$, here x was RGB image, y_1

[1] See [7] for more examples. Any task that has a fixed operator is a candidate.

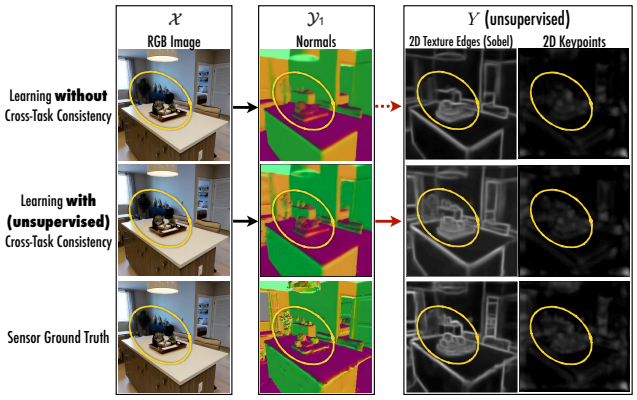


Figure 1: **Learning with cross-task consistency with *unsupervised domains* shown for a sample query**. The conventions of the figure is the same as Fig. 4 of the main paper. Using the notation $\mathcal{X}\rightarrow\mathcal{Y}_1\rightarrow Y$, here $\mathcal{X}$ is RGB image, $\mathcal{Y}_1$ is surface normals, and two domains in Y are 2D texture edges (Sobel filter) and 2D keypoints., depth, and occlusion edges. Learning with Cross-Task Consistency constraints still improved the results, even though no supervision was used in the introduced consistency constraints.

was surface normals, and the two domains in Y were 2D edges and 2D keypoints. The loss was $\mathcal{L}^{perceptual}_{\mathcal{X}\mathcal{Y}_1 Y}$.

The unsupervised consistency domains improved certain relevant features in the predicted normals. We expect the improvement to increase with more tasks added, though the gain will plateau beyond a certain point as the tasks will start to become redundant. Thus we expect the consistency tasks would need to be 'engineered' beyond a certain point to squeeze out further improvements.

Generally speaking, the usefulness of unsupervised tasks extends the applicability of the proposed method, in terms of improving the fit to the data, to single/few task datasets.

4. Handling of Ill-Posed Tasks

As noted in the introduction and Section 3.1.2 of the main paper, the task sets we consider may not be *informationally equivalent* – in the sense in that, in theory, one may not always be able to convert one to the other without some uncertainty. For example, we can expect to derive normals from depth, but the opposite direction is ill-posed. Therefore employing such an ill-posed link $normal\rightarrow depth$ as a loss for training a task like $RGB\rightarrow normal$ (in other words the triangle $RGB\rightarrow normal\rightarrow depth$) may appear problematic. That would be because, in theory it is impossible to infer the correct depth even given ground truth normals. Here we describe three points toward resolving this issue:

I) Most links are not ill-posed to learn, despite analytical definitions: Even though a link may be ill-posed in theory, the contextual information visible in the datapoint often resolves the uncertainty; just like the fact that when humans look at ground truth normals, they can infer the depth of the same scene based on the semantics and contextual information visible in the normal image. We found that the trained neural networks were surprisingly good at using this knowledge. This is well presented in Fig.4-(lower row) of the main paper, where the normals could be well predicted out of intuitively surprising and ill-posed tasks (e.g. see $TextureEdges\rightarrow normal$ or $3DCurvature\rightarrow normal$ or $SurfKeypoints\rightarrow normal$). Hence most of the links in the complete graph of tasks turn out to be not significantly ill-posed, despite their analytical definition. This also underscores the advantages of a fully computational and data-driven method vs those that primarily rely on analytical relationships [4], as many of the cross-task links we successfully use would be analytically ill-posed.

II) The perceptual loss formulation handles the residual error: Even if a link is not ill-posed, we still learn them using neural networks, which likely results in a small but non-zero loss after convergence. This means the link $\mathcal{Y}_1 \xrightarrow{f_{\mathcal{Y}_1\mathcal{Y}_2}} \mathcal{Y}_2$ will *not* yield the perfect $\mathcal{Y}_2^*$ even if provided with perfect $\mathcal{Y}_1^*$ in the input, i.e. $f_{\mathcal{Y}_1\mathcal{Y}_2}(\mathcal{Y}_1^*) \neq \mathcal{Y}_2^*$. Thus, there will always be residual imperfections in the link, and consequently when trying to employ it to optimize the task $\mathcal{X} \xrightarrow{f_{\mathcal{X}\mathcal{Y}_1}} \mathcal{Y}_1$ using the triangle $\mathcal{X} \xrightarrow{f_{\mathcal{X}\mathcal{Y}_1}} \mathcal{Y}_1 \xrightarrow{f_{\mathcal{Y}_1\mathcal{Y}_2}} \mathcal{Y}_2$, the imperfections of $\mathcal{Y}_1 \xrightarrow{f_{\mathcal{Y}_1\mathcal{Y}_2}} \mathcal{Y}_2$ may corrupt the link $\mathcal{X} \xrightarrow{f_{\mathcal{X}\mathcal{Y}_1}} \mathcal{Y}_1$.[2] (Again, that will be due to the fact that even outputting perfect $\mathcal{Y}_1^*$ by $f_{\mathcal{X}\mathcal{Y}_1}(\mathcal{X})$ would not result in estimating $\mathcal{Y}_2^*$ in the end of the triangle $f_{\mathcal{Y}_1\mathcal{Y}_2} \circ f_{\mathcal{X}\mathcal{Y}_1}(\mathcal{X})$ since $f_{\mathcal{Y}_1\mathcal{Y}_2}(\mathcal{Y}_1^*) \neq \mathcal{Y}_2^*$.) Therefore using a loss of the form $\mathcal{L} = \|f_{\mathcal{Y}_1\mathcal{Y}_2} \circ f_{\mathcal{X}\mathcal{Y}_1}(\mathcal{X}) - \mathcal{Y}_2^*\|$ in the framework to train $\mathcal{X} \xrightarrow{f_{\mathcal{X}\mathcal{Y}_1}} \mathcal{Y}_1$ would be problematic and corrupt the network.

We handle this by adopting a *perceptual loss* [3] based formulation. This basically means we use the loss

$$\mathcal{L} = \|f_{\mathcal{Y}_1\mathcal{Y}_2} \circ f_{\mathcal{X}\mathcal{Y}_1}(\mathcal{X}) - f_{\mathcal{Y}_1\mathcal{Y}_2}(\mathcal{Y}_1^*)\|, \qquad (1)$$

in lieu of $\mathcal{L} = \|f_{\mathcal{Y}_1\mathcal{Y}_2} \circ f_{\mathcal{X}\mathcal{Y}_1}(\mathcal{X}) - \mathcal{Y}_2^*\|$. This simple trick enforces that predicting perfect $\mathcal{X}_2^*$ by $f_{\mathcal{X}\mathcal{Y}_1}(\mathcal{X})$ would result in exactly loss 0 in the triangle loss, hence the residual imperfections of $\mathcal{Y}_1 \xrightarrow{f_{\mathcal{X}\mathcal{Y}_1}} \mathcal{Y}_2$ would not propagate to learning of $\mathcal{X} \xrightarrow{f_{\mathcal{Y}_1\mathcal{Y}_2}} \mathcal{Y}_1$ via the triangle loss $\mathcal{X} \xrightarrow{f_{\mathcal{X}\mathcal{Y}_1}} \mathcal{Y}_1 \xrightarrow{f_{\mathcal{Y}_1\mathcal{Y}_2}} \mathcal{Y}_2$.

In addition, the above perceptual loss based formulation makes training datasets with pair annotations $(\mathcal{X}, \mathcal{Y}_1)$ & $(\mathcal{Y}_1, \mathcal{Y}_2)$, rather than triplet $(\mathcal{X}, \mathcal{Y}_1, \mathcal{Y}_2)$ or higher order annotations, sufficient for learning with the cross-task consistency triangle $\mathcal{X} \xrightarrow{f_{\mathcal{X}\mathcal{Y}_1}} \mathcal{Y}_1 \xrightarrow{f_{\mathcal{Y}_1\mathcal{Y}_2}} \mathcal{Y}_2$. That is because the

[2] Note that $\mathcal{X} \xrightarrow{f_{\mathcal{X}\mathcal{Y}_1}} \mathcal{Y}_1 \xrightarrow{f_{\mathcal{Y}_1\mathcal{Y}_2}} \mathcal{Y}_2$ is the same as $\mathcal{X} \longrightarrow \mathcal{Y}_1 \longrightarrow \mathcal{Y}_2$ in the rest of the paper, just annotated with functions for ease of followed notations.

ground truth for the third domain $\mathcal{Y}_2^*$ is not used in the above perceptual loss in Eq. 1, hence at no point all three domains $(\mathcal{X}, \mathcal{Y}_1, \mathcal{Y}_2)$ need be *simultaneously* known for one datapoint.

III) Removing the low-mutual-information edges: If a link between two domains is severely problematic to the extent that the points **I** and **II** do not address it, we simply remove them from the complete task graph. We do that by performing vanilla pre-training of links between all pairs of domains $(\mathcal{X}_a \times \mathcal{X}_b)$ and removing those that do not results in a sufficiently low loss. However, we barely faced such links (see point **I** above) and noticed removing or not removing them did not make a notable difference in the overall results after optimizing the entire system.

5. Balancing Different Loss Terms

As discussed in the main paper, the final (total) loss is:

$$\sum_{i=1}^{N} \left(\left| f_{\mathcal{X}\mathcal{Y}_i}(x) - y_i \right| + \lambda \sum_{j=1}^{N} \left| f_{\mathcal{Y}_i\mathcal{Y}_j} \circ f_{\mathcal{X}\mathcal{Y}_i}(x) - f_{\mathcal{Y}_i\mathcal{Y}_j}(y_i) \right| \right),$$

which we can rewrite as:

$$\sum_{i=1}^{N} \left(\mathcal{L}^{\text{direct}} + \lambda \sum_{j=1}^{N} \mathcal{L}^{\text{percept}}_{\mathcal{X}\mathcal{Y}_i\mathcal{Y}_j} \right),$$

One issue is that if the number of perceptual losses is large, the perceptual losses will come to dominate the loss. We found that choosing λ to compensate for this effect improved the quality of the final networks.

We set λ per-batch and per-loss in order to make the magnitude of the perceptual losses' total gradient contribution independent of the number of perceptual losses used. Specifically, we set λ as follows:

$$\lambda_{ij}(x) = \frac{|\nabla \mathcal{L}^{\text{direct}}| + \sum_{k \neq i,j} |\nabla \mathcal{L}^{\text{percept}}_{\mathcal{X}\mathcal{Y}_i\mathcal{Y}_j}|}{(N-1)\left(|\nabla \mathcal{L}^{\text{direct}}| + \sum_k |\nabla \mathcal{L}^{\text{percept}}_{\mathcal{X}\mathcal{Y}_i\mathcal{Y}_j}|\right)},$$

where $|\cdot|$ is the ℓ^1 norm.

6. Optimizing the standard direct loss does not lead to optimizing cross-task losses

In this section we experimentally demonstrate that optimizing the standard direct loss (e.g. MSE loss) for a certain task does not naturally lead to optimizing the related cross-task losses; even if the direct loss appears to go down. In other words, optimizing for $\mathcal{X}\rightarrow\mathcal{Y}_1$ does not substantially optimize the red cross-task loss in $\mathcal{X}\rightarrow\mathcal{Y}_1\rightarrow\mathcal{Y}_2$.

This is demonstrated in Fig. 2, where various cross-task losses (i.e. the '$\rightarrow$'s) are plotted for a network that is being trained only for the standard direct MSE loss (orange curve) and one that is being trained with the standard direct MSE loss as well as cross-task losses (red curve). For both cases, the optimization starts from a baseline MSE-only network that is halfway to its convergence. As apparent in the figure, the network being optimized with only the direct loss does not reduce the perceptual losses, despite the fact that the direct loss was being successfully optimized till full convergence. This is while the global minimum of both the direct loss ($\mathcal{X}\rightarrow\mathcal{Y}_1$) and the cross-task losses ($\mathcal{X}\rightarrow\mathcal{Y}_1\rightarrow\mathcal{Y}_2$) are the same (predicting the ground truth for $\mathcal{Y}_1$). This again echoes that, given the existing sub-optimal optimizers, the process of training neural networks benefits from explicit augmentation of cross-task consistency based losses. This is inline with the quantitative and qualitative results provided in the main paper.

The results were reported using (re)Shading task. The conclusion was the same for other tasks.

7. Derivation of Generic Consistency Criterion

In the main paper we derived the triangle consistency constraint. The derivation for the general path-consistency constraint is similar.

Here's the setup: say that we have two paths ($P = P_1 ... P_{k-1} P_\ell$ and $Q = Q_1 ... Q_{\ell-1} Q_\ell$). We want to train the first edge in the path P, $f_{\mathcal{X}P_1}$. Assume this edge does not also appear in Q. Note that we are requiring $P_1 = Q_1 = \mathcal{X}$ and $P_k = Q_\ell$. We can then define the path composition function:

$$f_P(x) \triangleq f_{P_{k-2}P_{k-1}} \circ f_{P_{k-1}P_k} \circ ... \circ f_{\mathcal{X}P_1}(x),$$

and the analogous functions: f_Q for Q, and $f_{P_{1:k'}}$ for sub-paths[3] $P_{1:k'}$ of P. The associated consistency constraint is:

$$|f_P(x) - f_Q(x)|$$

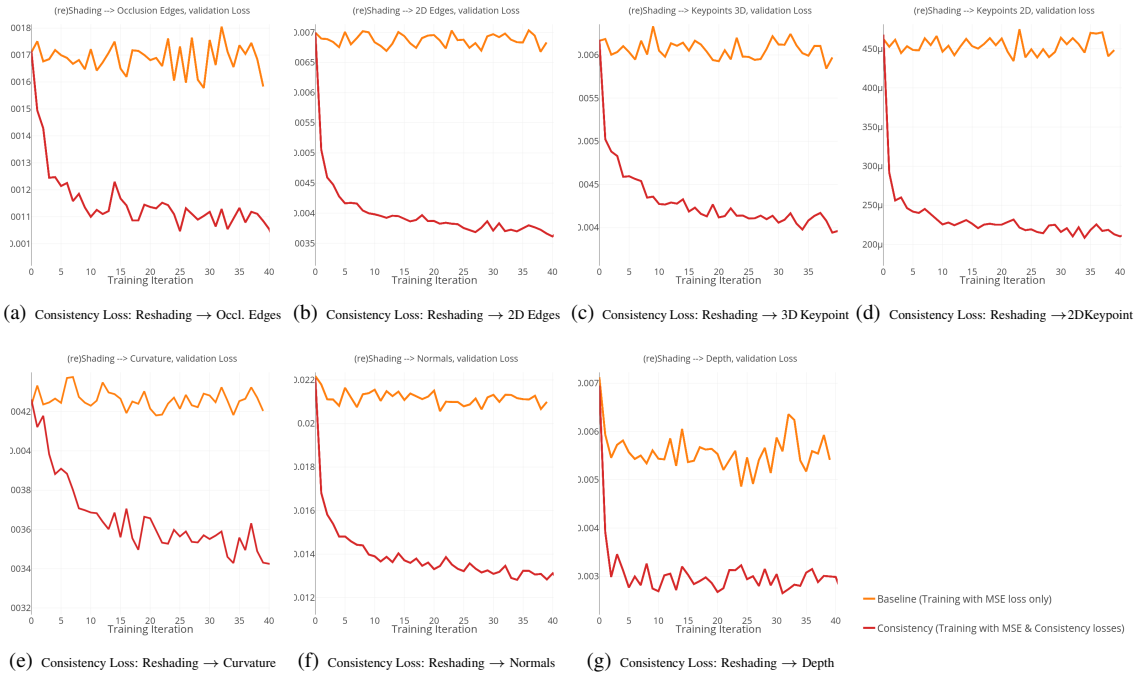
.

The derivation requires mapped triplets $(x, p_{k-1}, p_k) \sim \mathcal{X} \times P_{k-1} \times P_k)$. In contrast, the final equation (and, therefore, training) only requires paired data $(x, p_{k-1}) \sim \mathcal{X} \times P_{k-1}$.

Proof. In the derivation, we'll absorb terms that are constant (w.r.t. the optimization parameters) into some catch-all, C.

$$\begin{aligned}
&|f_P(x) - f_Q(x)| \\
&\quad \leq |f_P(x) - p_k| + |p_k - f_Q(x)| \\
&\quad = |f_P(x) - p_k| + C \\
&\quad \leq |f_{P_{k-1}P_k} \circ f_{P_{1:k-1}}(x) - f_{P_{k-1}P_k}(p_{k-1})| \\
&\qquad\qquad + |f_{P_{k-1}P_k}(p_{k-1}) - p_k| + C \\
&\quad = |f_{P_{k-1}P_k} \circ f_{P_{1:k-1}}(x) - f_{P_{k-1}P_k}(p_{k-1})| + C
\end{aligned}$$

[3] $P_{1:k'}$ is the same as P up until and including $P_{k'}$ for $k' \leq k$.



(a) Consistency Loss: Reshading → Occl. Edges (b) Consistency Loss: Reshading → 2D Edges (c) Consistency Loss: Reshading → 3D Keypoint (d) Consistency Loss: Reshading →2DKeypoint

(e) Consistency Loss: Reshading → Curvature (f) Consistency Loss: Reshading → Normals (g) Consistency Loss: Reshading → Depth

Figure 2: **Optimizing the standard direct loss does not lead to optimizing cross-task losses.** Various cross-task losses are plotted for a network that is being trained only for the standard direct MSE loss (orange curve) and one that is being trained with the standard direct MSE loss as well as cross-task losses (red curve). The network being optimized with only the direct loss does not reduce the perceptual losses, despite the fact that the direct loss was being successfully optimized till full convergence. This echos the necessity of augmentation the training process with explicit losses based on cross-task consistencies.

□

8. Sensitivity Analysis: Edge Selection

As described in Sec.3.2 of the main paper, multiple possibilities exists for selecting the path to be optimized at each iteration. We adopted selecting the most inconsistent (*Maximally Violating*) path at the time. We provide a discussion and experimental justification here.

Table 1 compares different strategies for selecting the path to optimize in message passing (i.e. *SelectNetwork* in Algorithm 1). *Random* Path Selection represents selecting one of the feasible paths $\mathcal{X}\rightarrow\mathcal{Y}_1\rightarrow\mathcal{Y}_2$ randomly. *Maximally Violating* Path represents selecting the most inconsistent path at the time. That is akin to optimizing to reduce the upper bound of inconsistency in the system. *All* Paths represents randomly picking a $\mathcal{X}\rightarrow\mathcal{Y}_1$ then using *all* of the perceptual losses that start from $\mathcal{Y}_1$, with the total loss being the sum of all of them. Consequently, this method is slow and has a large memory requirement as all many networks are in use at the same time.

As mentioned in the main paper, the results in Table 1 shows picking the *Maximally Violating* Path gave the best results, though there was not a significance difference among the methods in terms of the final performance. In terms of the computation and memory requirements, using *all* paths is significantly more expensive. This signifies the potential value in adopting a proper selection criterion in message passing. We used *Maximally Violating* selection method in our experiments. The amount of violation/inconsistency of a path $\mathcal{X}\rightarrow\mathcal{Y}_1\rightarrow\mathcal{Y}_2$ is defined to be the loss of that path at the predicting $\mathcal{Y}_2$ given $\mathcal{X}$, normalized by a fixed constant which was found by via grid search.

Path selection method	MSE (↓)
Maximally Violating Path	**1.78**
Random Path Selection	1.83
All Paths	1.88

Table 1: Comparison of different path selection strategies in message passing.

Table 1 is reported using (re)Shading task. The conclusions remain the same for different tasks.

9. Sensitivity Analysis: Path Lengths

The proposed method is applicable to optimizing cross-task consistency with any inferences path lengths. By path length, we are referring to the the number of cascaded edges in $\mathcal{X}\rightarrow\mathcal{Y}_1\rightarrow\mathcal{Y}_2\rightarrow\mathcal{Y}_3\rightarrow\dots$. Here we provide a discussion

and experiments on this aspect. Table 2 compares the final results of optimizing with paths with maximum length 3 vs maximum length 2.

Method	MSE
Path length 3 ($\mathcal{X}\to\mathcal{Y}_1\to\mathcal{Y}_2\to\mathcal{Y}_3$)	1.80
Path length 2 ($\mathcal{X}\to\mathcal{Y}_1\to\mathcal{Y}_3$)	1.78

Table 2: Impact of different path lengths in optimization.

As described in the main paper, there was a negligible difference associated with using longer paths in our experiments. We believe this is due to the fact that our task graph is nearly complete, hence for any path $\mathcal{X}\to\mathcal{Y}_1\to\mathcal{Y}_2\to\mathcal{Y}_3$ with length 3, the corresponding path with length 2 with the same end points, i.e. $\mathcal{X}\to\mathcal{Y}_1\to\mathcal{Y}_3$, is also included in the optimization. Hence the cross-task consistency value of a path with length 3 is already enforced by a path with length 2, as far as optimizing the link $\mathcal{X}\to\mathcal{Y}_1$ is concerned. The same discussion applies to paths longer than 3 as well. Hence, as described in Sec. 3.2, we limited the length of sampled paths to 2 in our experiments as shorter paths are computationally cheaper. However, the proposed framework is applicable to arbitrary lengths, especially if the task graph is not complete.

Table 2 is reported using (re)Shading task. The conclusion was the same for other tasks.

10. Standard Error Over Multiple Seeds

The tables below add *Standard Error* columns to the tables 1-3 of the main paper. The Standard Error is reported over multiple (3) independent runs ("**Std. Err. (3 Runs)**") of our network and the main baseline as well as across test set images ("**Std. Err. (Test Set)**"). The low Standard Error denotes statistical significance in the observed trends. Also, please note that the reported improvement margins (e.g. 0.87 vs 1.00) are quite significant (13% relative improvement) and geometrically noteworthy.

11. Results on NYUv2 Dataset

We used NYU v2 [5] dataset in addition to Taskonomy [7] and Replica [6] for evaluations. Table 4 shows our method and the main baseline, denoting that trends remain the same. Note that NYU v2 was used for evaluation only, so the training dataset remained Taskonomy. These results show the evaluation trends reported in the main paper were not specific to Taskonomy dataset. They also suggest Taskonomy provided a better training data for achieving good geometric prediction results compared to NYU v2 (as training our baseline UNet on Taskonomy and testing on NYU v2 reported better results compared to training on NYU v2 and testing on NYU v2– see Table 1 of [4]) possibly due to its larger size and type of sensor ground truth. Hence we adopted Taskonomy as the training dataset for a more reliable experimentation.

12. More Metrics

We provide additional evaluations on the Taskonomy and Replica datasets here in the supplementary material to show that the results are not specific to the ℓ^1-norm. Specifically we provide additional task-specific and perceptual metrics that were omitted, for space, from Table 1 in the main paper. The trends and conclusions remain the same as in the main paper.

12.1. Additional Task-Specific Metrics for Direct Prediction

Some tasks (surface normal estimation, depth estimation) also have additional common standard metrics beyond MSE or ℓ^1. Common task-specific metrics evaluating models on direct prediction of RGB to X are provided for

- **Surface normals**: Angular error metrics are provided for Taskonomy (Table 10) and Replica (Table 7), and NYUv2 (Table 4).
- **Depth estimation**: Depth error metrics are provided for Taskonomy (Table 8) and Replica (Table 5).

For reShading, since there are not commonly-accepted metrics, we provide both L1 and MSE error.

Common angular error metrics: The Tables 10 & 7 show five metrics beyond L1 and MSE: mean and median of angular surface normal error, and the proportion of predictions within 11.25°, 22.5°, and 30° of the ground-truth. The trends and conclusions remain the same as in the main paper.

Common depth error metrics: The Tables 8 & 5 show the same results as the direct L1 prediction from Table 1 in the main paper, with the addition of standard depth estimation metrics: Relative Error (Rel. Err), Scale Invariant Logarithmic error (SILog, main benchmark for KITTI [2]), Inverse Root Mean Squared Error (IRMSE, RMSE is used for NYUv2 [5]) and Logarithmic Error ($\log_{10}$). The trends and conclusions remain the same as in the main paper.

12.2. Additional Perceptual Metrics

We also provide additional perceptual metrics for both Replica and Taskonomy. Fig. 6 shows results using additional perceptual tasks (*Keypoints 2D, Keypoints 3D*, and *Edges 3D*) and metrics (MSE) to those shown in Table 1 on Taskonomy in the main paper. In Replica, there are currently only the three provided tasks, so we show MSE loss in addition to L1 from the main paper. The trends and conclusions remain the same as in the main paper.

Surface Normal Estimation Method	MSE	**Std. Err. (Test Set)**	**Std. Err. (3 Runs)**
Baseline (UNet with MSE loss)	1.00	**0.0037**	**0.0059**
Consistency (UNet with consistency loss)	0.87	**0.0052**	**0.0050**

Method (Depth)	Error MSE	Error **Std. Err. (Test)**
Baseline (UNet MSE)	0.40	**0.0006**
Consistency	0.36	**0.0007**

Method (Reshading)	Error MSE	Error **Std. Err. (Test)**
Baseline (UNet MSE)	2.20	**0.0005**
Consistency	1.78	**0.0003**

Table 3: **Tables 1-3 of main paper with addition of *Standard Error* values (in bold)**. Results show statistical significance in observed trends, as the standard errors indicate that these results are extremely unlikely to occur from chance.

(Surface Normal Est.) Method	Error (↓) Mean°	Med.°	L1	MSE	Accuracy (↑) ≤11.25°	22.5°	30°
Baseline (UNet)	**9.91**	**6.16**	**8.65**	**2.08**	**0.71**	**0.88**	**0.93**
Cycle Consistency	12.83	9.85	11.01	2.63	0.57	0.83	0.91
GeoNet (Impr.)	10.27	6.47	8.97	2.22	0.70	0.87	0.92
GeoNet (Orig.)	**10.08**	6.49	**8.79**	**2.09**	**0.71**	**0.88**	**0.93**
Multitask	11.63	7.50	10.15	2.69	0.66	0.84	0.91
Pix2Pix	13.07	10.04	11.29	2.72	0.57	0.84	0.91
Prcpt. Loss (ImageNet)	11.06	7.14	9.65	2.53	0.68	0.85	0.91
Prcpt. Loss (Random)	11.29	7.45	9.85	2.57	0.66	0.85	0.91
Consistency	**9.88**	**6.06**	**8.62**	**2.11**	**0.71**	**0.88**	**0.93**

Table 4: **Evaluating surface normal prediction on NYU v2 test set.** Values **bolded and starred*** indicate the best-performing method. Values that are **bolded** but not starred indicate methods that were statistically indistinguishable from the best-performing method (2-sample paired t-test, $\alpha = 0.01$).

In general, evaluation of high dimensional regression tasks, in comparison to lower dimensional classification, is more challenging as often no single metric captures all the desirable properties of a prediction. Thus using multiple metrics simultaneously provide a more complete evaluation picture for such cases.

13. More Qualitative Results

We provide some randomly selected (non-cherry picked) sample images to give a sense of the general performance of the consistency-trained and the baseline networks.

13.1. Queries From the Taskonomy Test Set

Fig. 8 shows the results on various test images sampled from two of the buildings in Taskonomy's test set. The buildings were picked randomly. Each row shows a query and the columns show the predictions, the ground truth, and error maps, for different prediction domains. **This figure is the same as Fig. 4-middle row of the main paper, but for more queries.**

13.2. Surface Normals From Middle Domains

Fig. 9 (without consistency) and Fig. 10 (with consistency) provide more results in the same format of Figure 3 in the main paper, with addition of error images and ground truth.

14. Blind Guess (Statistically Informed Guesses)

As described in the paper, we compared all networks against a "statistically informed guess". This is a guess in that it does not look at the input x when predicting the label y. This is statistically informed in that it is the best-possible such guess given the "guess" constraint. Specifically, we compute the guess, g^* as

$$g^* \triangleq \arg\min_g E_y[|g - y|]$$

.

A visualization of these guesses for the *normals*, *reshading*, and *depth* are provided in Fig. 7.

15. Code, Examples, and Docker

We've open-sourced our code and are providing tools for training and evaluating models using consistency. The repository is available here, and contains (among other things):

- Pretrained models
- Demo code
- Uncertainty energy estimation code
- Training scripts
- Docker and installation instructions

Please see the README in that repository for the most up-to-date information.

References

[1] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006. 1

[2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 5

[3] Justin Johnson, Alexandre Alahi, and Fei-Fei Li. Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155, 2016. 2

[4] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018. 2, 5

Replica Direct Prediction (Additional Metrics)

(Depth Estimation) Method	Error (↓) L1	MSE	IRMSE	Log10	SI Log
Baseline UNet	1.99	0.07	**0.00***	0.13	**0.01***
Constant Pred.	4.81	0.38	0.00	0.32	0.07
GeoNet (Impr.)	1.83	0.06	**0.00**	0.13	**0.01**
GeoNet (Orig.)	4.01	0.30	0.00	0.23	0.04
Multitask	2.44	0.11	0.00	0.16	0.02
Taskonomy	3.72	0.24	0.00	0.23	0.02
Consistency	**1.63***	**0.05***	**0.00**	**0.12***	**0.01**

Table 5: **Depth estimation on Replica.**

(reShading) Method	Error (↓) L1	MSE
Baseline UNet	**9.55**	**1.77**
Constant Pred.	16.45	4.35
Multitask	10.32	1.95
Taskonomy	11.43	2.27
Consistency	**9.22***	**1.69***

Table 6: **ReShading estimation on Replica.**

(Surface Normal Est.) Method	Error (↓) Mean°	Median°	L1	MSE	Accuracy (↑) ≤11.25°	22.5°	30°
Baseline UNet	5.76	**2.63**	4.96	1.02	0.87	0.95	0.97
Constant Pred.	19.13	16.27	16.02	4.94	0.36	0.65	0.77
Cycle Consistency	8.36	4.84	7.13	1.54	0.77	0.91	0.95
GeoNet (Impr.)	**5.46***	**2.59**	**4.70***	**0.95***	**0.88***	**0.95***	**0.97***
GeoNet (Orig.)	11.50	7.19	7.48	1.98	0.68	0.86	0.91
Multitask	7.02	3.51	6.03	1.45	0.84	0.92	0.95
Pix2Pix	9.03	5.95	7.70	1.60	0.75	0.91	0.95
Prcpt. Loss (ImageNet)	**5.62**	**2.57***	4.85	**0.99**	0.87	**0.95**	**0.97**
Prcpt. Loss (Random)	5.78	2.74	4.99	1.00	0.86	**0.95**	**0.97**
Taskonomy	19.13	16.27	16.02	4.94	0.36	0.65	0.77
Consistency	**5.60**	**2.63**	**4.80**	**0.99**	0.88	**0.95**	**0.97**
0.25% Data: Baseline (UNet)	8.83	4.13	7.61	2.27	0.78	**0.89**	**0.91**
0.25% Data: Consistency	**8.46**	**3.77**	**7.28**	**2.05**	**0.79**	**0.89**	**0.92**

Table 7: **Surface normal estimation on Replica.**

Figure 3: **Extended quantitative results on Replica.** Values **bolded and starred*** indicate the best-performing method. Values that are **bolded** but not starred indicate methods that were statistically indistinguishable from the best-performing method (2-sample paired t-test, $\alpha = 0.01$)

Taskonomy Direct Prediction (Additional Metrics)

(Depth Estimation) Method	Error (↓) L1	MSE	IRMSE	Log_{10}	Rel. Err.	SI Log
Baseline (UNet)	**2.27**	**0.17**	**0.00**	**0.13**	**0.17**	**0.02**
Constant Pred.	7.07	0.87	0.00	0.41	0.63	0.07
GeoNet (Impr.)	**2.26***	**0.16***	**0.00***	**0.13***	**0.17***	**0.02**
GeoNet (Orig.)	4.07	0.44	0.00	0.22	0.31	0.05
Multitask	2.81	0.20	0.00	0.17	0.22	0.02
Taskonomy	4.55	0.44	0.00	0.26	0.28	0.03
Consistency	**2.29**	**0.16**	**0.00**	0.13	**0.17**	**0.02***

Table 8: **Depth estimation on Taskonomy.**

(reShading) Method	Error (↓) L1	MSE
Baseline (UNet)	**10.45***	**3.21**
Constant Pred.	24.85	8.91
Multitask	11.61	3.36
Taskonomy	16.58	5.36
Consistency	**10.52**	**3.21***

Table 9: **ReShading estimation on Taskonomy.**

(Surface Normal Est.) Method	Error (↓) Mean°	Median°	L1	MSE	Accuracy (↑) ≤11.25°	22.5°	30°
Baseline (UNet)	**6.86**	**2.42**	**5.95**	**1.58**	**0.81**	**0.91**	**0.94**
Constant Pred.	21.06	19.14	17.80	5.73	0.27	0.56	0.76
Cycle Consistency	10.09	6.29	8.68	2.06	0.69	0.87	0.93
GeoNet (Impr.)	**6.81***	**2.37***	**5.91***	**1.57***	**0.81***	**0.91***	**0.94***
GeoNet (Orig.)	15.49	11.38	9.58	2.86	0.51	0.76	0.86
Multitask	8.17	3.66	7.07	1.84	0.78	0.89	0.93
Pix2Pix	10.92	7.28	9.40	2.26	0.68	0.87	0.92
Prcpt. Loss (ImageNet)	6.98	2.50	6.06	1.62	0.81	0.91	0.94
Prcpt. Loss (Random)	7.10	2.68	6.17	1.62	0.81	0.91	0.94
Taskonomy	8.70	4.21	7.54	1.96	0.76	0.89	0.93
Consistency	7.01	2.52	6.08	1.63	0.81	0.91	0.94
0.25% Data: Baseline	**9.43**	**4.05**	**8.17**	**2.43**	**0.74**	**0.86**	**0.91**
0.25% Data: Consistency	10.63	5.17	9.19	2.78	0.70	0.84	0.89

Table 10: **Surface normal estimation on Taskonomy.**

Figure 4: **Extended quantitative results on the Taskonomy test set.** Values **bolded and starred*** indicate the best-performing method. Values that are **bolded** but not starred indicate methods that were statistically indistinguishable from the best-performing method (2-sample paired t-test, $\alpha = 0.01$)

[5] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. *Indoor Segmentation and Support Inference from RGBD Images*, pages 746–760. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. 5

[6] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 5

[7] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018. 1, 5

Replica Perceptual Results (Extended)

Depth Est. →	Surface Normal		reShading	
Method	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)
Baseline UNet	10.47	3.08	12.99	3.12
Constant Pred.	22.23	7.80	19.94	6.00
GeoNet (Impr.)	10.47	3.07	12.75	3.03
GeoNet (Orig.)	13.88	5.03	14.03	4.30
Multitask	15.30	5.33	16.14	4.56
Taskonomy	18.06	6.04	15.39	3.86
Consistency	**7.01***	**1.71***	**11.21***	**2.57***

Table 11: **Perceptual results for depth estimation on Replica.**

reShading →	Surface Normal		Depth Estimation	
Method	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)
Baseline UNet	6.90	1.59	2.74	0.12
Constant Pred.	15.74	5.13	5.14	0.41
Multitask	7.24	1.81	3.36	0.18
Taskonomy	8.70	2.41	3.85	0.23
Consistency	**5.50***	**1.18***	**1.96***	**0.07***

Table 12: **Perceptual results for reShading on Replica.**

Surface Normal →	Depth Estimation		reShading	
Method	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)
Baseline UNet	4.69	0.37	13.15	2.80
Constant Pred.	4.75	0.43	33.31	15.15
Cycle Consistency	5.65	0.54	22.39	7.30
GeoNet (Impr.)	4.62	0.36	12.79	2.65
GeoNet (Orig.)	6.23	0.81	19.34	7.24
Multitask	5.58	0.53	22.11	7.20
Pix2Pix	4.52	0.39	19.03	5.87
Prcpt. Loss (ImageNet)	3.45	0.20	**8.31***	**1.43***
Prcpt. Loss (Random)	4.88	0.41	15.34	3.61
Taskonomy	3.73	0.26	33.31	6.62
Consistency	**2.07***	**0.08***	9.99	1.69
0.25% Data: Baseline	5.65	0.53	21.76	7.03
0.25% Data: Consistency	**2.41**	**0.10**	**12.26**	**2.69**

Table 13: **Perceptual results for surface normal estimation on Replica.**

Figure 5: **Quantitative perceptual results for surface normal estimation on Replica.** Values **bolded and starred*** indicate the best-performing method. Values that are **bolded** but not starred indicate methods that were statistically indistinguishable from the best-performing method (2-sample paired t-test, $\alpha = 0.001$)

Taskonomy Perceptual Results (Extended)

Depth Estimation →	Surface Normal		reShading		Princ. Curvature		Keypts. (2D, SURF)		Keypts (3D, NARF)		Edges (2D, Sobel)		Edges (3D, Occ.)	
Method	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)
Baseline UNet	13.62	4.33	15.68	5.28	7.31	2.19	3.50	0.36	7.56	1.10	12.61	3.49	**1.46***	**0.12**
Constant Pred.	22.37	8.31	27.27	10.72	7.96	2.29	3.85	0.42	7.88	1.56	12.77	4.82	1.97	0.16
GeoNet (Impr.)	13.77	4.41	15.76	5.27	7.52	2.26	3.49	0.36	7.69	1.12	12.67	3.50	**1.46**	**0.12***
GeoNet (Orig.)	15.44	5.94	18.73	7.63	4.03	1.04	**2.66***	**0.27***	6.56	1.09	10.78	**2.76***	2.18	0.22
Multitask	17.18	6.12	19.55	7.14	7.54	2.28	3.39	0.35	9.55	1.53	13.67	3.68	1.91	0.13
Taskonomy	18.82	6.40	20.83	7.17	6.65	1.75	3.44	0.36	9.72	1.48	14.10	4.16	1.94	0.14
Consistency	**9.46***	**2.56***	**12.66***	**4.06***	**3.61***	**0.85***	3.55	0.35	**5.69***	**0.88***	**9.82***	3.00	1.52	0.13

Table 14: **Perceptual results for depth estimation on Taskonomy.**

reShading →	Surface Normal		Depth Estimation		Princ. Curvature		Keypts. (2D, SURF)		Keypts (3D, NARF)		Edges (2D, Sobel)		Edges (3D, Occ.)	
Method	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)
Baseline UNet	9.58	2.59	3.38	0.25	3.78	0.92	2.98	0.28	6.22	1.06	10.85	3.50	1.53	0.15
Constant Pred.	19.96	7.40	7.14	0.88	3.53	0.91	3.17	0.33	7.48	1.67	12.62	5.53	1.83	0.17
Multitask	9.19	2.48	3.54	0.27	3.56	0.86	2.96	0.27	6.23	1.10	10.75	3.30	1.53	0.15
Taskonomy	11.72	3.45	4.69	0.49	3.54	0.90	3.09	0.31	6.93	1.44	11.19	4.26	1.60	0.16
Consistency	**7.13***	**1.88***	**2.51***	**0.18***	**3.28***	**0.79***	**2.93***	**0.24***	**5.40***	**0.83***	**9.38***	**2.85***	**1.35***	**0.12***

Table 15: **Perceptual results for reShading on Taskonomy.**

Surface Normal →	Depth Estimation		reShading		Princ. Curvature		Keypts. (2D, SURF)		Keypts (3D, NARF)		Edges (2D, Sobel)		Edges (3D, Occ.)	
Method	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)	L1 (↓)	MSE (↓)
Baseline UNet	8.17	1.21	20.94	8.09	3.41	0.84	3.91	0.40	6.48	1.06	9.98	3.38	1.52	0.16
Cycle Consistency	8.81	1.44	30.33	14.83	3.84	0.92	3.88	0.40	7.53	1.18	10.26	3.11	1.69	0.17
GeoNet (Impr.)	8.18	1.21	20.84	8.05	3.40	0.83	3.91	0.40	6.43	1.06	9.99	3.40	1.52	0.16
GeoNet (Orig.)	7.71	1.46	27.35	15.27	**3.32**	0.81	2.97	0.31	7.64	1.30	**9.09***	**2.51***	1.48	0.15
Multitask	8.78	1.41	27.32	12.82	3.65	0.91	3.94	0.41	7.21	1.12	10.16	3.38	1.64	0.17
Pix2Pix	8.12	1.27	26.23	11.23	3.83	0.93	3.92	0.40	7.80	1.21	10.33	3.39	1.75	0.17
Prcpt. Loss (ImageNet)	6.86	0.88	17.36	5.70	3.36	0.80	3.77	0.38	6.01	0.98	9.63	3.08	1.45	0.14
Prcpt. Loss (Random)	8.59	1.36	23.98	10.26	3.41	0.83	3.91	0.40	6.75	1.10	10.01	3.40	1.56	0.16
Consistency	**4.32***	**0.36***	**12.15***	**3.38***	**3.29***	**0.76***	**2.94***	**0.24***	**5.48***	**0.89***	9.50	2.89	**1.36***	**0.12***
0.25% Data: Baseline (UNet)	8.86	1.42	26.91	12.33	3.78	0.97	3.95	0.41	7.16	1.14	10.31	3.54	1.60	0.16
0.25% Data: Consistency	**5.07**	**0.50**	**15.96**	**5.01**	**3.74**	**0.90**	**3.77**	**0.38**	**6.35**	**1.04**	**9.93**	**2.97**	**1.57**	**0.15**

Table 16: **Perceptual results for surface normal estimation on Taskonomy.**

Figure 6: **Extended quantitative perceptual results on the Taskonomy test set.** Values **bolded and starred*** indicate the best-performing method. Values that are **bolded** but not starred indicate methods that were statistically indistinguishable from the best-performing method (2-sample paired t-test, $\alpha = 0.001$). MSE is shown in addition to L1, and results are shown for additional tasks (*Keypoints 2D, Keypoints 3D*, and *Edges 3D*).

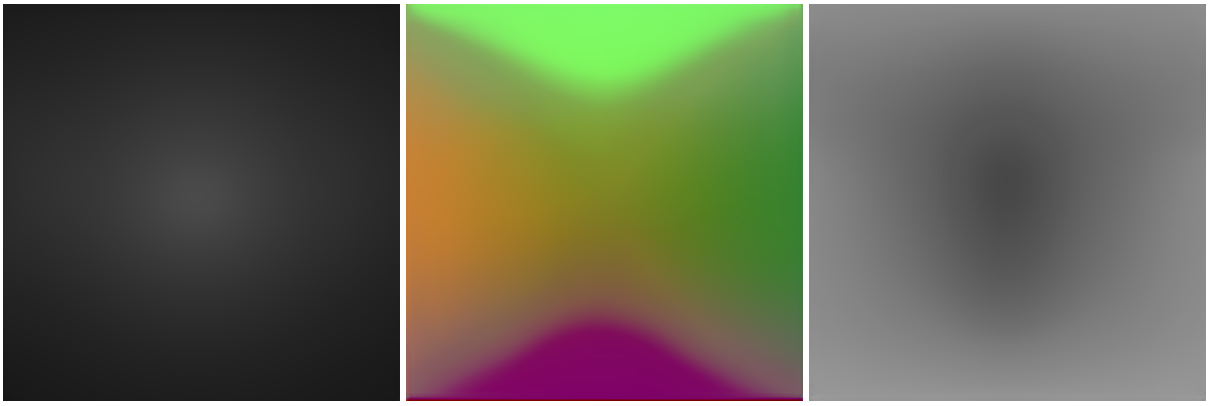


Figure 7: **Statistically informed guesses ("Blind Guess") on the Taskonomy dataset.** Left: Depth. Middle: Surface Normals. Right: (re)Shading. These images minimize expected L1 error on the training dataset: $\min_g E_y[|g - y|]$.

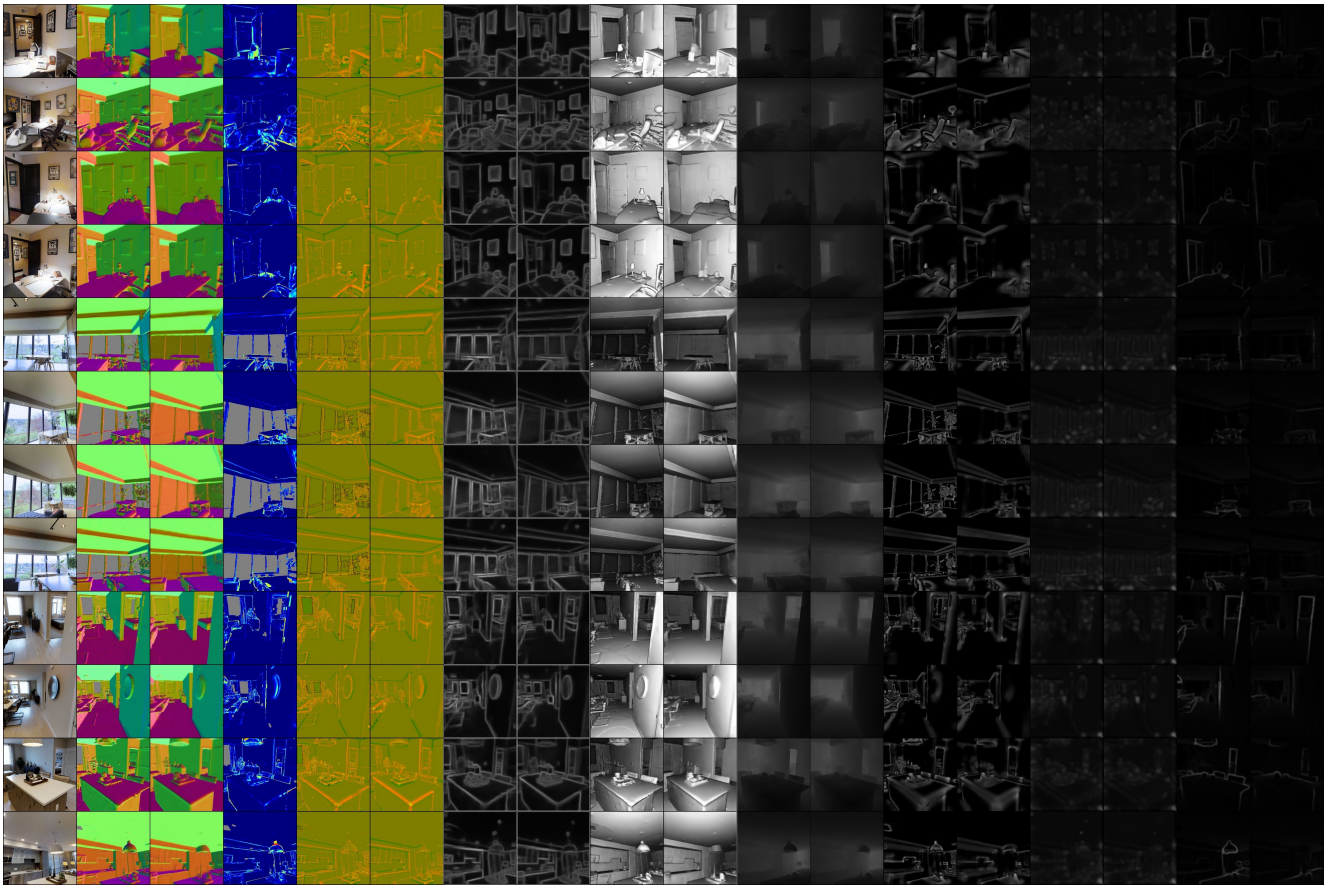


Figure 8: More qualitative results on test set images of Taskonomy dataset. **This figure is the same as Fig. 4-middle row of the main paper, but for more queries.** The first four columns show: query, surface normal prediction, surface normal ground truth, and error map. Other domains show prediction and ground-truth. The domains from left to right are: surface normals, 3D curvature, Sobel edges, reshading, depth, 3D keypoints, 2D keypoints, occlusion edges. [best seen on screen & zoomed in]

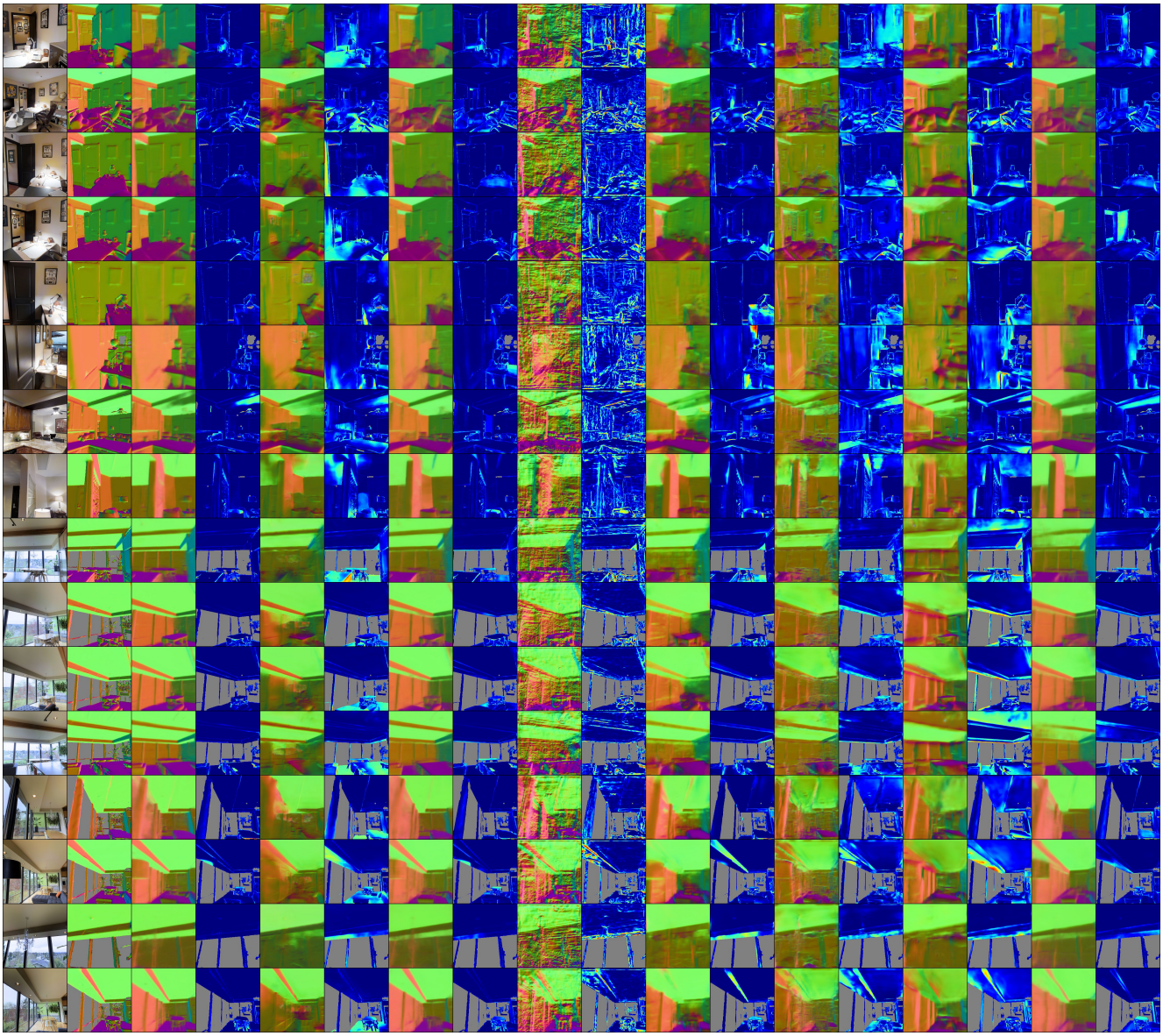


Figure 9: More results on estimating surface normal out of middle domains without enforcing consistency. **This figure is the same as Fig. 3-upper row of the main paper, but for more queries**. The ground truth and RGB image are shown on the left. The error map of each prediction is show on its right. [best seen on screen & zoomed in]

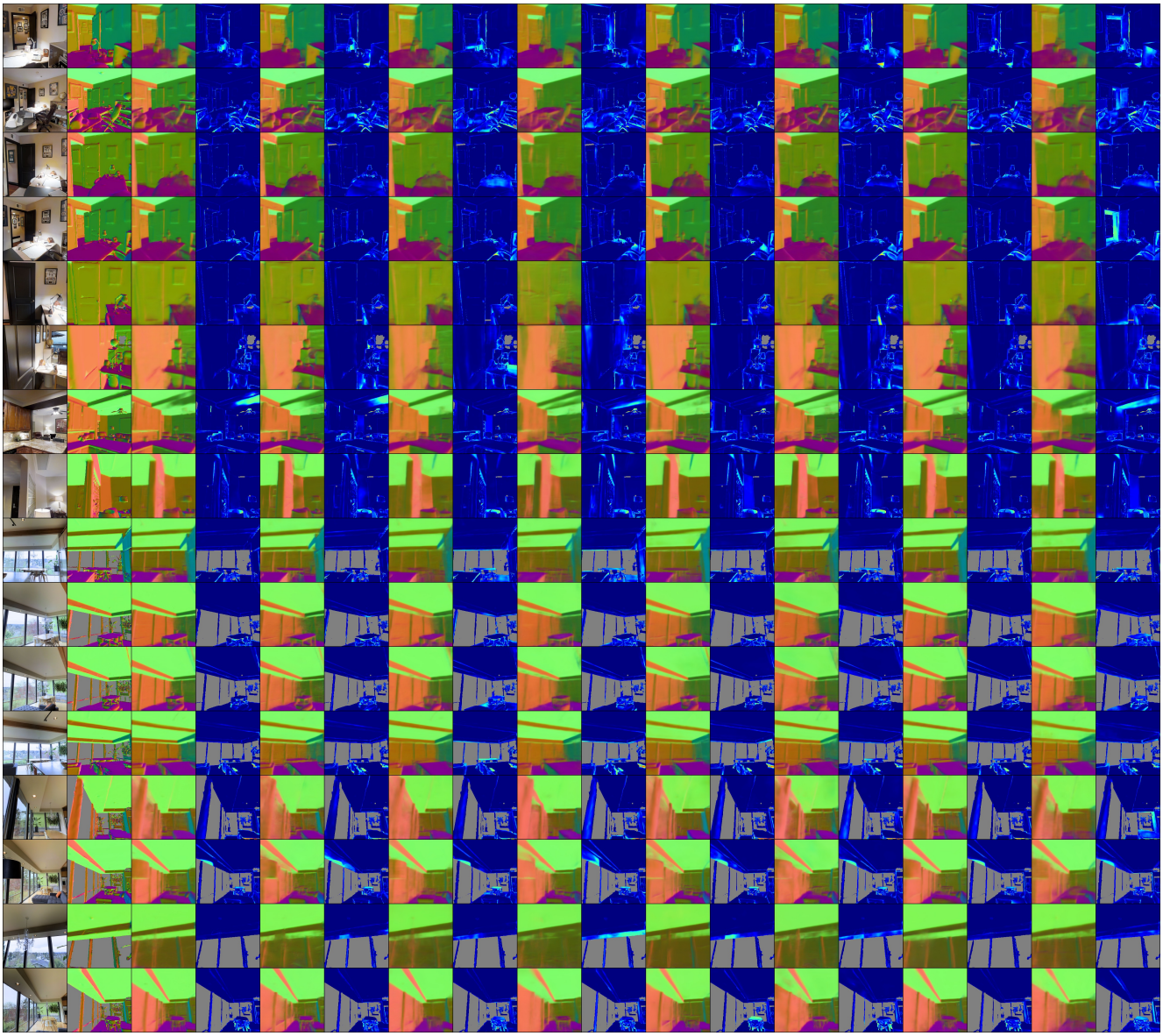


Figure 10: More results on estimating surface normal out of middle domains after enforcing consistency. **This figure is the same as Fig. 3-lower row of the main paper, but for more queries**. The ground truth and RGB image are shown on the left. The error map of each prediction is show on its right. [best seen on screen & zoomed in]