TOM-Net: Learning Transparent Object Matting from a Single Image

Guanying Chen, Kai Han, Kwan-Yee K. Wong
The University of Hong Kong, Hong Kong
{gychen, khan, kykwong}@cs.hku.hk

Abstract

This paper addresses the problem of transparent object matting. Existing image matting approaches for transparent objects often require tedious capturing procedures and long processing time, which limit their practical use. In this paper, we first formulate transparent object matting as a refractive flow estimation problem. We then propose a deep learning framework, called TOM-Net, for learning the refractive flow. Our framework comprises two parts, namely a multi-scale encoder-decoder network for producing a coarse prediction, and a residual network for refinement. At test time, TOM-Net takes a single image as input, and outputs a matte (consisting of an object mask, an attenuation mask and a refractive flow field) in a fast feed-forward pass. As no off-the-shelf dataset is available for transparent object matting, we create a large-scale synthetic dataset consisting of 178K images of transparent objects rendered in front of images sampled from the Microsoft COCO dataset. We also collect a real dataset consisting of 876 samples using 14 transparent objects and 60 background images. Promising experimental results have been achieved on both synthetic and real data, which clearly demonstrate the effectiveness of our approach.

1. Introduction

Image matting refers to the process of extracting the foreground matte of an image by locating the region of the foreground object and estimating the opacity of each pixel inside the foreground region. The foreground object can then be composited onto a new background image using the matting equation

\[ C = F + (1 - \alpha)B, \quad \alpha \in [0, 1], \] (1)

where \( C \) denotes the composited color, \( F \) the foreground color, \( B \) the background color, and \( \alpha \) the opacity.

Image matting has been widely used in image editing and film production. However, most of the existing methods are tailored for opaque objects, and cannot handle transparent objects whose appearances depend on how light is refracted from the background.

To model the effect of refraction, Zongker et al. [26] introduced environment matting as

\[ C = F + (1 - \alpha)B + \Phi, \quad \alpha \in [0, 1], \] (2)

where \( \Phi \) is the contribution of environment light caused by refraction or reflection at the foreground object. Besides estimating the foreground shape, environment matting also describes how objects interact with the background.

Many efforts [3, 22, 15, 25, 7, 5] have been devoted to improving the seminal work of [26]. The resulting methods often require either a huge number of input images to achieve a higher accuracy, or specially designed patterns to reduce the number of required images. They are in general all very computational expensive.

In this paper, we focus on environment matting for transparent objects. It is highly ill-posed, if not impossible, to estimate an accurate environment matte for transparent objects from a single image with an arbitrary background. Given the huge solution space, there may exist multiple objects and backgrounds that can produce the same refractive effect. In order to make the problem more tractable, we simplify our problem to estimating an environment matte that...
can produce visually realistic refractive effect from a single image, instead of estimating a highly accurate refractive flow. We define the environment matte in our model as a triple consisting of an object mask, an attenuation mask and a refractive flow field. Realistic refractive effect can then be obtained by compositing the transparent object onto new background images (see Fig. 1).

Inspired by the great successes of convolutional neural networks (CNNs) in high-level computer vision tasks, we propose a convolutional neural network, called TOM-Net, for simultaneous learning of an object mask, an attenuation mask and a refractive flow field from a single image with an arbitrary background. The key contributions of this paper can be summarized as follows:

- We introduce a simple and efficient model for transparent object matting as simultaneous estimation of an object mask, an attenuation mask and a refractive flow field.
- We propose a convolutional neural network, TOM-Net, to learn an environment matte of a transparent object from a single image. To the best of our knowledge, TOM-Net is the first CNN that is capable of learning transparent object matting.
- We create a large-scale synthetic dataset and a real dataset as a benchmark for learning transparent object matting. Our TOM-Net has produced promising results on both the synthetic and real datasets.

2. Related Work

In this section, we briefly review representative works on environment matting and recent works on CNN based image matting.

Environment matting Zongker et al. [26] introduced the concept of environment matting, and assumed each foreground pixel being originated from a single rectangular region of the background. They obtained the environment matte by identifying the corresponding background region for each foreground pixel using three monitors and multiple images. Chuang et al. [3] extended [26] in two ways. First, they replaced the single rectangular supporting area for a foreground pixel with multiple 2D oriented Gaussian strips. This makes it possible for their method to model the effects of color dispersion, multiple mapping and glossy reflection. Second, they simplified the environment matting equation by assuming the object colorless and specular transparent. This allows them to achieve real time environment matting (RTEm). The environment matte was then extracted with one image taken in front of a pre-designed pattern. However, RTEm requires background images to segment the transparent objects, and depends on a time-consuming off-line processing. Wexler et al. [22] introduced a probabilistic model based method which assumes each background point has a probability to make contribution towards the color of a certain foreground point. Their approach does not require pre-designed patterns during data acquisition, but it still needs multiple images and can only model thin transparent objects. Peers and Dutré [15] used a large number of wavelet basis backgrounds to obtain the environment matte, and their method can also model the effect of diffuse reflection. Based on the fact that a signal can be decomposed uniquely in the frequency domain, Zhu and Yang [25] proposed a frequency-based approach to extract an accurate environment matte. They used Fourier analysis to solve the decomposition problem. Both [15] and [25] require a large number of images to extract the matte (e.g., [15] needs 2,400 images and [25] needs 4,096 images for an image of size $1024 \times 1024$), making them not very practical. Recently, compressive sensing theory has been applied to environment matting to reduce the number of images required. Duan et al. [6] applied this theory in the spatial domain and Qian et al. [16] applied it in the frequency domain. However, the number of images needed is still in the order of hundreds. In contrast, our work can estimate an environment matte from a single image in a fast feed-forward computation without the need for pre-designed patterns or additional background images.

Yeung et al. [24] proposed an interactive way to estimate an environment matte given an image containing a transparent object. Their method requires users to manually mark the foreground and background in the image, and models the refractive effect using a thin-plate-spline transformation. Their method does not produce an accurate environment matte, but instead a visually pleasing refractive effect. Our method shares the same spirit, but does not involve any human interaction.

CNN based image matting Although the potential of CNN on transparent object matting has not yet been explored, some existing works have adopted CNNs for solving the general image matting problem. Shen et al. [18] introduced a CNN for image matting of color portrait images. Cho et al. [2] proposed a network to predict a better alpha matte by taking the matting results of the traditional method and normalized color image as input. Xu et al. [25] introduced a deep learning framework that can estimate an alpha matte from an image and its trimap. However, none of these methods can be applied directly to the task of transparent object matting as object opacity alone is not sufficient to model the refractive effect.

3. Matting Formulation

As a transparent object may have multiple optical properties (e.g., color attenuation, translucency and reflection), estimating an accurate environment matte for a generic transparent object from a single image is very challenging. Following the work of [3], we cast environment matting to a refractive flow estimation problem by assuming that each
foreground pixel only originates from one point in the background due to refraction. Compared to the seminal work of [26], which models each foreground pixel as a linear combination of a patch in the background, our formulation is more tractable and can be easily encoded using a CNN.

In [26], the per-pixel environment matting is obtained through leveraging color information from multiple background images. Given a set of pre-designed background patterns, matting is formulated as

$$C = F + (1 - \alpha)B + \sum_{i=1}^{m} R_i M(T_i, A_i),$$  \(\text{(3)}\)

where \(F\), \(B\) and \(\alpha\) denote the foreground color, background color and weight, respectively. The last term in (3) accounts for the environment light accumulated from \(m\) pre-designed background images (\(m = 3\) in [26]). \(R_i\) is a factor describing the contribution of light emanating from the \(i\)-th background image \(T_i\), \(M(T_i, A_i)\) denotes the average color of a rectangular region \(A_i\) on the background image \(T_i\).

To obtain an environment matte, the transparent object is placed in front of the monitor(s), and multiple pictures of the object are captured with the monitor(s) displaying different background patterns\(^1\). Generally, a surface point receives light from multiple directions, especially for a diffuse surface. When it comes to a (perfectly) transparent object, however, a surface point will only receive light from one direction as determined by the law of refraction. Consider a single background image as the only light source, the problem can be modeled as

$$C = F + (1 - \alpha)B + R M(T, P),$$  \(\text{(4)}\)

where \(M(T, P)\) is a bilinear sampling operation at location \(P\) on the background image \(T\). Further, by assuming a colorless transparent object, we have \(F = 0\) and \(R\) becomes a light attenuation index \(\rho\) (a scalar value). The formulation in (4) can be simplified to

$$C = (1 - \alpha)B + \rho M(T, P),$$  \(\text{(5)}\)

where \(\rho \in [0, 1]\) denotes the attenuation index.

Here, we use refractive flow to model the refractive effect of a transparent object. The refractive flow of a foreground pixel is defined as the offset between the foreground pixel and its refraction correspondence on the background image.

We further introduce a binary foreground mask to define the object region in the image. The matting equation can now be rewritten as

$$C = (1 - m)B + m \rho M(T, P),$$  \(\text{(6)}\)

where \(m \in \{0, 1\}\) denotes background \((m = 0)\) or foreground \((m = 1)\). The matte can then be estimated by solving \(m\), \(\rho\) and \(P\) for each pixel in the input image containing the transparent object\(^2\).

4. Learning Transparent Object Matting

In this section, we present a two-stage deep learning framework, called TOM-Net, for learning transparent object matting (see Fig. 2). The first stage, denoted as CoarseNet, is a multi-scale encoder-decoder network that takes a single image as input, and predicts an object mask, an attenuation mask and a refractive flow field simultaneously. CoarseNet is capable of predicting a robust object mask. However, the estimated attenuation mask and refractive flow field lack local structural details. To overcome this problem, we introduce the second stage of TOM-Net, denoted as RefineNet, to achieve a sharper attenuation mask and a more detailed refractive flow field. RefineNet is a residual network [10] that takes both the input image and the output of CoarseNet as input. After training, our TOM-Net can predict an environment matte from a single image in a fast feed-forward pass.

\(^1\)For an image of size 512 × 512, 18 pictures and around 20 minutes processing time are needed.

\(^2\)For each observed pixel, we have 7 unknowns (3 for \(B\), 2 for \(P\), 1 for \(m\) and 1 for \(\rho\)).
4.1. Encoder-Decoder for Coarse Prediction

The first stage of our TOM-Net (i.e., CoarseNet) is based on mirror-link CNN introduced in [19]. Mirror-link CNN was proposed to learn non-lambertian object intrinsic decomposition. Its output consists of an albedo map, a shading map and a specular map. It shares a similar output structure with our transparent object matting task (i.e., three output branches sharing the same spatial dimensionality). Therefore, it is reasonable for us to adapt mirror-link CNN for our CoarseNet.

The mirror-link CNN adapted for our CoarseNet consists of one shared encoder and three distinct decoders. The encoder contains six down-sampling convolutional blocks, leading to a down-sampling factor of 64 in the bottleneck layer. Features in the encoder layers are connected to the decoder layers having the same spatial dimensions through skip connections [17]. Cross-links [19] are introduced to make different decoders share the same input in each layer, so that decoders can better utilize the correlation between different predictions.

Learning with multi-scale loss has been proved to be helpful in dense prediction tasks (e.g., [8, 9]). Since we formulate the problem of transparent object matting as refractive flow estimation, which is a dense prediction task, we augment our mirror-link CNN with multi-scale loss similar to [9]. We use four different scales in our model, where the first scale starts from the decoder features with a down-sampling factor of 8 and the largest scale has the same spatial dimensions as the input.

In contrast to the recent two stage framework for image matting [23], our TOM-Net has a shared encoder and three parallel decoders to accommodate different outputs. Also, we augment our CoarseNet with multi-scale loss and cross-link. Moreover, TOM-Net is trained from scratch while the encoder in [23] is initialized with the pre-trained VGG16.

4.2. Loss Function for Coarse Stage

CoarseNet takes a single image as input and predicts the environment matte as a triple consisting of an object mask, an attenuation mask and a refractive flow field. The learning of CoarseNet is supervised by the ground-truth matte using an object mask segmentation loss $L_{ms}$, attenuation regression loss $L_{ar}$, and refractive flow regression loss $L_{fr}$.

Besides, the predicted matte is expected to render an image as close to the input image as possible when applied to the ground-truth background. Hence, in addition to the supervision of the matte, we also take image reconstruction loss $L_{ir}$ into account. Note that the ground-truth background is only used to calculate the reconstruction error during training but not needed during testing. CoarseNet can therefore be trained by minimizing

$$
\mathcal{L}^c = \alpha_{ms}^c L_{ms} + \alpha_{ar}^c L_{ar} + \alpha_{fr}^c L_{fr} + \alpha_{ir}^c L_{ir},
$$

where $\alpha_{ms}^c, \alpha_{ar}^c, \alpha_{fr}^c, \alpha_{ir}^c$ are weights for the corresponding loss terms.

Object mask segmentation loss Object mask segmentation is simply a spatial binary classification problem. The output of the object mask decoder has a dimension of $2 \times H \times W$, where $H$ and $W$ denote the height and width of the input. We normalize the output with $softmax$ and compute the loss using the binary cross-entropy function

$$
L_{ms} = - \frac{1}{HW} \sum_{ij} (\tilde{M}_{ij} \log(P_{ij}) + (1 - \tilde{M}_{ij}) \log(1 - P_{ij})),
$$

where $\tilde{M}_{ij} \in \{0, 1\}$ and $P_{ij} \in [0, 1]$ represent ground truth and normalized foreground probability of the pixel at $(i, j)$, respectively.

Attenuation regression loss The predicted attenuation mask has a dimension of $1 \times H \times W$. The value of this mask is in the range of $[0, 1]$, where $0$ indicates no light can pass and $1$ indicates the light will not be attenuated. We adopt a mean square error (MSE) loss

$$
L_{ar} = \frac{1}{HW} \sum_{ij} (A_{ij} - \tilde{A}_{ij})^2,
$$

where $A_{ij}$ is the predicted attenuation index and $\tilde{A}_{ij}$ the ground truth at $(i, j)$.

Refractive flow regression loss The predicted refractive flow field has a dimension of $2 \times H \times W$, where we have one channel for the horizontal displacement and another for the vertical displacement. We normalize the refractive flow with $tanh$ activation and multiply it by the width of the input, such that the output is constrained in the range of $[-W, W]$. We adopt an average end-point error (EPE) loss

$$
L_{fr} = \frac{1}{HW} \sum_{ij} \sqrt{(F^x_{ij} - \tilde{F}^x_{ij})^2 + (F^y_{ij} - \tilde{F}^y_{ij})^2},
$$

where $(F^x, F^y)$ and $(\tilde{F}^x, \tilde{F}^y)$ denote the predicted flow and the ground truth, respectively.

Image reconstruction loss We use MSE loss to measure the dissimilarity between the reconstructed image and the input image. Denoting the reconstructed image by $\tilde{I}$ and the ground-truth image (i.e., the input image) by $I$, the reconstruction loss is given by

$$
L_{ir} = \frac{1}{HW} \sum_{ij} \|I_{ij} - \tilde{I}_{ij}\|_2^2.
$$

Implementation details In all experiments, we empirically set $\alpha_{ms}^c = 0.1, \alpha_{ar}^c = 1, \alpha_{fr}^c = 0.01$, and $\alpha_{ir}^c = 1$. The loss weights for different scales are $\frac{1}{2^s}$, where $s \in \{1, 2, 3, 4\}$ denotes the scale. CoarseNet contains $8M$ parameters and it takes about 2.5 days to train with Adam optimizer [12] on a single GPU. We first train the CoarseNet from scratch until convergence and then train the RefineNet.
4.3. Residual Learning for Matte Refinement

As the attenuation mask and the refractive flow field predicted by the CoarseNet lack structural details, a refinement stage is needed to produce a detailed matte. Observing that residual learning is particularly suitable for tasks whose input and output are largely similar [11, 14], we propose a residual network, denoted as RefineNet, to refine the matte predicted by the CoarseNet.

We concatenate the input image and the output of the CoarseNet to form the input of the RefineNet. As the object mask predicted by the CoarseNet is already plausible, the RefineNet only outputs an attenuation mask and a refractive flow field. The parameters of the CoarseNet are fixed when training the refinement stage.

**Loss for the refinement stage** The overall loss for the refinement stage is

\[
L' = \alpha_{ar} L_{ar} + \alpha_{fr} L_{fr},
\]

where \( L_{ar} \) is the refinement attenuation regression loss, \( L_{fr} \) the refinement flow regression loss, and \( \alpha_{ar}, \alpha_{fr} \) their weights. The definitions of these two losses are identical to those defined in the first stage. We found that adding the image reconstruction loss in the refinement stage did reduce the image reconstruction error during training, but was not helpful in reserving sharp edges of the refractive flow field (e.g., mouth of a glass), which is essential to maintain the details of an object. Therefore, we remove the image reconstruction loss here. The \( \tanh \) activation for refractive flow is also omitted in this stage to encourage the network to focus on boundary regions that may have larger prediction errors.

**Implementation details** We set \( \alpha_{ar} = 1, \alpha_{fr} = 1 \) for the refinement. RefineNet contains 1.1M parameters and it takes about 2 days to train with Adam optimizer. RefineNet is randomly initialized during training.

5. Dataset for Learning and Evaluation

As no off-the-shelf dataset for transparent object mating is available, and it is very tedious and difficult to produce a large training set with ground-truth object masks, attenuation masks and refractive flow fields, we created a large-scale synthetic dataset by using \textit{POV-Ray} [1] to render images of synthetic transparent objects. Besides, we also collected a real dataset for evaluation. We will show that our TOM-Net trained on the synthetic dataset can generalize well to real world objects, demonstrating its good transferability.

5.1. Synthetic Dataset

We used a large number of background images and 3D models to render our training samples. We randomly changed the pose of the models, as well as the viewpoint and focal length of the camera in the rendering process to avoid overfitting to a fixed setting.

**Backgrounds Images** We employed two types of background images, namely scene images and synthetic patterns. For scene images, we randomly sampled images from the Microsoft COCO [13] dataset\(^3\). The background images for the training set are sampled from COCO Train2014 and Test2015, while that for the validation set are from COCO Val2014, giving rise to 100K scene images in total. For synthetic patterns, we rendered 40K patterns of size \(512 \times 512\) using \textit{POV-Ray} built-in textures.

**Transparent Objects** We divided common transparent objects into four categories, namely glass, glass with water, lens, and complex shape (see Fig. 4 for examples). We constructed parametric 3D models for the first three categories, and generated a large number of models using random parameters. For complex shapes, we constructed parametric 3D models for basic shapes like sweeping-spheres and squashed surface of revolution (SOR) parts, and composed a larger number of models using these basic shapes. We generated 178K 3D models in total, with each model assigned a random refractive index \( \lambda \in [1.3, 1.5] \). The distribution of these models in four categories is shown in Tab. 1 (first two rows).

<table>
<thead>
<tr>
<th>Type</th>
<th>Glass</th>
<th>Glass &amp; Water</th>
<th>Lens</th>
<th>Complex</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic Train</td>
<td>52K</td>
<td>26K</td>
<td>20K</td>
<td>80K</td>
<td>178K</td>
</tr>
<tr>
<td>Synthetic Val</td>
<td>250</td>
<td>250</td>
<td>200</td>
<td>200</td>
<td>900</td>
</tr>
<tr>
<td>Real Test</td>
<td>470</td>
<td>103</td>
<td>61</td>
<td>242</td>
<td>876</td>
</tr>
</tbody>
</table>

**Ground-truth Matte Generation** We obtained the ground-truth object mask of a model by rendering it in front of a black background image and setting its color to white. Similarly, we obtained the ground-truth attenuation mask of a model by simply rendering it in front of a white background image. Finally, we obtained the ground-truth refractive flow field (see Fig. 4) of a model by rendering it in front of a sequence of Gray-coded patterns.

**Data Augmentation** To improve the diversity of the training data and narrow the gap between real and synthetic data, extensive data augmentation was carried out on-the-fly. For an image with a size of \(512 \times 512\), we randomly performed color (brightness, contrast and saturation) augmentation (in a range of \([-0.2, 0.2]\)), image scaling (in a range of \([0.875, 1.05]\)), noise perturbation (in a range of \([-0.05, 0.05]\)), and horizontal/vertical flipping. Besides, we also blurred the object boundary to make the synthetic data visually more natural. A patch with a size of \(448 \times 448\) was then randomly cropped from an augmented image and used as input.

\(^3\)Other large-scale datasets like ImageNet [4] can also be used.
Table 2. Ablation study results. F, A, I, and M are short for flow, attenuation, image reconstruction, and object mask, respectively. (The first value for EPE is measured on the whole image and the second measured within the object region. A-MSE and I-MSE are computed on the whole image.)

<table>
<thead>
<tr>
<th>ID</th>
<th>Model Variants</th>
<th>F-EPE (m)</th>
<th>A-MSE (%)</th>
<th>I-MSE (%)</th>
<th>M-IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Background</td>
<td>6.5 / 41.0</td>
<td>1.58</td>
<td>0.87</td>
<td>0.15</td>
</tr>
<tr>
<td>1</td>
<td>CoarseNet - (C_r)</td>
<td>3.9 / 26.5</td>
<td>0.24</td>
<td>0.23</td>
<td>0.98</td>
</tr>
<tr>
<td>2</td>
<td>CoarseNet - (C_f)</td>
<td>2.3 / 15.7</td>
<td>0.25</td>
<td>0.22</td>
<td>0.98</td>
</tr>
<tr>
<td>3</td>
<td>CoarseNet - (multi-scale)</td>
<td>2.4 / 16.6</td>
<td>0.69</td>
<td>0.25</td>
<td>0.94</td>
</tr>
<tr>
<td>4</td>
<td>RefineNet - (cross-link)</td>
<td>2.5 / 17.2</td>
<td>0.30</td>
<td>0.31</td>
<td>0.97</td>
</tr>
<tr>
<td>5</td>
<td>CoarseNet</td>
<td>2.2 / 15.4</td>
<td>0.25</td>
<td>0.18</td>
<td>0.97</td>
</tr>
<tr>
<td>6</td>
<td>CoarseNet + RefineNet</td>
<td>2.0 / 13.7</td>
<td>0.25</td>
<td>0.19</td>
<td>0.97</td>
</tr>
</tbody>
</table>

To validate the transferability of TOM-Net, we introduce a real dataset, which was collected using 14 objects and 60 background images, resulting in a dataset of 876 images. The data distribution is summarized in Table 1 (last row). During the data capturing process, the objects were placed under different poses, with the distances between the camera, object and background uncontrolled. Fig. 5 (second column) shows some sample images from the real dataset. Note that we do not have the ground-truth matte for the real dataset. We instead captured images of the backgrounds without the transparent objects to facilitate evaluation.

6. Experiments and Results

In this section, we present experimental results and analysis. Currently, it is non-trivial to have a fair comparison with the previous methods, since none of the them can compute the matte from a single image of a transparent object, and there exists no common datasets and measurements for evaluation. We performed ablation study for TOM-Net, and evaluated our approach on both synthetic and real data. In addition, a user study was conducted to validate the realism of TOM-Net composites. Our code, model and datasets will be made available online: https://guanyingc.github.io/TOM-Net.

6.1. Ablation Study for Network Structure

We quantitatively analyzed different components of TOM-Net using synthetic dataset. In particular, we verified the effectiveness of image reconstruction loss ($L_c$), refractive flow regression loss ($L_f$), multi-scale loss, cross-link, and RefineNet, where the first four components were evaluated by removing each of them from CoarseNet. Each variant was trained separately. RefineNet was evaluated by adding it to a trained CoarseNet and was trained while fixing the parameters of CoarseNet. Besides, we included a naive baseline, denoted as Background, by considering a zero matte case (i.e., whole image as object mask, no attenuation, and no refractive flow) where the reconstructed image is the same as the background image. We evaluated endpoint error (EPE) for refractive flow fields, intersection over union (IoU) for object masks, mean square error (MSE) for attenuation masks and image reconstruction results, respectively. The results are summarized in Table 2. Background was outperformed by all TOM-Net variants with a large margin for all the evaluation metrics, which clearly shows that TOM-Net can successfully learn the matte. Removing each component from CoarseNet, the overall performance decreased, although some metrics slightly increased due to learning trade-offs, demonstrating these four components are essential for TOM-Net. By introducing RefineNet, the refractive flow performance was boosted, which verified the effectiveness of RefineNet (see Fig. 3).

6.2. Results on Synthetic Data

Quantitative results for synthetic validation dataset are presented in Tab. 3. We compared TOM-Net against Background and CoarseNet. Here, to accelerate training convergence, we first trained CoarseNet from scratch using our synthetic dataset excluding the complex shape subset. The trained CoarseNet was then fine-tuned using the entire training set including complex shapes, followed by training of RefineNet on the entire training set with random initialization. Similar to previous experiments, TOM-Net outperformed Background with a large margin, and slightly outperformed CoarseNet in EPE and MSE, which implies more local details can be learned by RefinedNet. The errors of complex shape category are larger than that of others, because complex shapes contain more sharp regions that will induce more errors. Although TOM-Net is not expected to learn highly accurate refractive flow, the average EPE er-
errors (2.7/18.6)\(^6\) are very small compared with the dimensionality of the input image (448 × 448). In this sense, our predicted flow is capable of producing visually plausible refractive effect (see Fig. 4). Although the background images and objects in the validation set never appear in the training set, TOM-Net can still predict robust matte. The pleasing results of the complex shapes also demonstrate that our model can generalize well from basic shapes to complex shapes.

### 6.3. Results on Real Data

We evaluated TOM-Net on our collected real dataset, which consists of 876 images of real objects. Due to the absence of ground-truth matte, evaluation on the absolute

---

\(^6\)The first value is measured on the whole image and the second measured within the object region.

---

Tables 3 and 4 present the quantitative results on synthetic and real data, respectively. Table 5 summarizes the user study results. These evaluations highlight the robustness and generalization capabilities of TOM-Net.
7. Conclusion and Discussion

We have introduced a simple and efficient model for transparent object matting, and proposed a CNN architecture, called TOM-Net, that takes a single image as input and predicts environment matte as an object mask, an attenuation mask, and a refractive flow field in a fast feed-forward pass. Besides, we created a large-scale synthetic dataset and a real dataset as a benchmark for learning transparent object matting. Promising results have been achieved on both synthetic and real data, which clearly demonstrated the feasibility and effectiveness of the proposed approach. Since our model assumes objects to be colorless and specular transparent, TOM-Net cannot be applied to colored transparent objects, translucent objects and transparent objects with multiple mapping (i.e., refraction and reflection happen simultaneously at a surface point). We consider exploring better models and architectures to handle these scenarios as our future work.

---

7glass × 12, glass & water × 4, lens × 2, and complex shape × 2.
Acknowledgments This project is supported by a grant from the Research Grant Council of the Hong Kong (SAR), China, under the project HKU 718113E. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. We thank Yiming Qian for help with the synthetic data rendering.

References